Grouped Goodness-of-Fit Tests for Binary Regression Models

by

Jana Dorthea Canary

BS (Mathematics), MS (Forest Ecology)

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

(Biostatistics)

University of Tasmania, December 2013

Supervisors	**Doctor Stephen Quinn**
	Associate Professor Leigh Blizzard
	Professor David Hosmer
Research Supervisor	**Professor Ronald Barry**

Dedication

To my parents Willa and Jim, my husband Mark, and my daughter Jacqueline

Declaration of Originality

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

Signed: …… Date: ……

Statement of Authorship

This thesis can be made available for loan. Copying of any part of this thesis is prohibited for two years from the date this statement was signed; after that time limited copying is permitted in accordance with the *Copyright Act 1968.*

Signed: …… Date: ……

Acknowledgements

I would like to first thank my primary supervisor, Dr. Stephen Quinn, for his support. I feel fortunate to have had Steve as a supervisor. He was always willing to take time to work with me, to give carefully thought out comments on my work, and to deal with the challenges of working with someone on the other side of the globe. I am very appreciative of his efforts in guiding me through the thesis process.

I would also like to thank my other supervisors. First, thank you to Dr. Leigh Blizzard, who made it possible for me to come to Menzies and helped me to navigate through the steps of setting up a Ph.D. candidature remotely from Alaska. He also provided useful comments on my work and provided financial support through an NHMRC grant that, along with an Australian Postgraduate Award, made my research possible. I also would like to thank Dr. Ron Barry of the University of Alaska, who was kind enough to meet with me to discuss my work while I was in Alaska, particularly during the early stages of my candidature. Lastly, I would like to thank Dr. David Hosmer, who is an emeritus professor and an expert in the field of biostatistics. He kindly offered his time and advice, and I very much appreciate his help.

I would also like to thank several fellow students and staff (in no particular order) at Menzies whose support both in person and electronically made the process so much easier: Kara, Laura, Dawn, Peta, Oliver, Kathy, Karen, Petr, Barbara, Tracey, Steve, David, and Ben. Thank you for your advice, support, encouragement, and friendship. Also my friends in Alaska and Hobart who provided childcare when I needed it, chats, and support: Petra, John, Gordon, Nancy, Jay, Donie, Peter, Margaret, Trusten, William, Leslie, and Sharon.

Also, I would like to thank the two anonymous examiners of my thesis. Their thorough review and thoughtful comments were very helpful and encouraging.

Lastly I would like to thank my husband Mark Conde and daughter Jacqueline. I cannot thank Mark enough for his financial and emotional support, and for lively discussions about mathematics. And thank you to Jacqueline, for the backrubs, hugs, and understanding when I did not have time to do fun things. I hope to make it up to you soon.

Abstract

How well a proposed regression model fits the observed outcome data is a critical question. The answer may influence model selection, and the conclusions drawn. Summary goodness-of-fit (GOF) statistics are used to assess model fit. Pearson's chi-squared GOF statistic $\left(X^2\right)$ is used to evaluate the fit of logistic regression models, but X^2 isn't appropriate when the model contains continuous covariates. Other GOF statistics are applicable, including the Hosmer-Lemeshow $\left(HL\right)$, Pigeon-Heyse $\left(J^2\right)$, and Tsiatis $\left(T\right)$ statistics. All have similarities to X^2 and group data artificially.

Simulation studies assessing new GOF statistics for logistic models with continuous covariates often include HL for comparison. We know of no study that compares HL, J^2, and T. We did so here, applying the same grouping method (deciles-of-risk) to all. Our results indicated that HL and T followed their reported distributions, but J^2 did not. Its distribution was closer to $J^2 \sim \chi^2\left(G-2\right)$, where *G*=groups, rather than the reported $\chi^2\left(G-1\right)$. Assuming $J^2 \sim \chi^2\left(G-2\right)$, T maintained the Type I error rate twice as often as HL and J^2. The rates of HL and J^2 were often lower than expected when dichotomous, quadratic, or interaction terms were included. The statistics had similar power to detect departures from a true underlying model.

The logistic model is the canonical generalized linear model (GLM) for binomial outcomes. Although many GOF statistics have been developed for logistic models, there are fewer for non-canonical GLM with binomial outcomes. The properties of the logistic model make the development of GOF statistics relatively straightforward, but it can be more difficult for non-canonical GLMs.

We considered whether HL, J^2, and T could be applied to non-canonical GLM with Bernoulli outcomes and continuous covariates. Our investigation found that HL and J^2 can be applied directly, but T cannot. We introduced an augmented version of the Tsiatis model and

generalised T, ($T_{\mathcal{G}}$). We showed that under non-canonical links, $T_{\mathcal{G}} \sim \chi^2(G)$. In a second simulation study, HL, J^2, and $T_{\mathcal{G}}$ were used to evaluate the fit of probit, log-log, complementary log-log and log binomial models. The deciles-of-risk method was applied. Type I error rates were consistently maintained by $T_{\mathcal{G}}$, while those of HL and J^2 were often lower than expected if the model included dichotomous, quadratic, or interaction terms. Because the distributions of HL and J^2 varied, it was unclear how their degrees-of-freedom could be adjusted. The statistics had similar power to detect an incorrect model in most situations. An exception occurred when a log model was incorrectly fit to data generated from a logistic model; here $T_{\mathcal{G}}$ had more power than HL or J^2.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

"*The oldest, shortest words – 'yes' and 'no' – are those which require the most thought.*"

– Pythagoras

Yes or no, success or failure, life or death - these are all examples of binary responses. Critical research questions can have answers that take this form. Once a scientist identifies such a research question, designs a study, toils for hours collecting data, and then builds a regression model appropriate for binary outcomes – after all of that, but before they can answer their research question, they must first answer another binary question: 'Does my model fit the data that I observed?' Their answer to this question will affect the conclusions they draw from all of their hard work. It is critical. Checking the fit of binary regression models is central to this thesis, and we began by giving a brief overview of the topic. Note that a more rigorous explanation of the material in this overview will be given in Chapters two and three. We conclude this introductory chapter with a list of the research questions that will be addressed in the thesis, and give a synopsis of the thesis presentation.

1.1 Background

So how can the researcher decide whether or not to keep their model? One way is to calculate a goodness-of-fit statistic that summarizes how well the outcomes predicted by their model fit the outcomes they observed. This was the basic idea behind Karl Pearson's well-known goodness-of-fit test (Pearson 1900). Pearson is considered to be the "nucleus of the movement of systematic statistical thinking"(Mukhopadhyay 2000). He was one of the first to introduce the concept of goodness-of-fit. Originally he considered a multinomial variate with a fully specified distribution function, whose range can be divided into some finite number of mutually exclusive classes. Then, since the distribution function is specified, the probability of any observation falling into a particular class can be calculated (Kendall, Stuart, Ord and Arnold 1999). One way to think of this is as multinomial trials where each observation is placed into one of a finite

number of cells in a contingency table. Under the null hypothesis, the theoretical multinomial probability of a being placed into a cell is equal to the probability derived from the observed data and the parameters specified under the null hypothesis. The alternative is that they are significantly different. Fisher extended Pearson's work by considering the case when the parameters are not fully specified (Kendall, et al. 1999). Later, Pearson's test was applied to regression models that can be fitted to obtain estimates of the parameters. This includes the logistic regression model, which is used in many fields to relate a categorical outcome variable to a set of predictor variables. It is a member of the class of models called *generalized linear models* (GLM), and can be used when the outcome data is assumed to come from a binomial distribution (McCullagh and Nelder 1989). If the outcome is binary, then it comes from the Bernoulli distribution, which is a special case of the binomial distribution. In regression models, the characteristics that correspond to the cells in the contingency table are the "covariate patterns". These are the unique combinations of the possible categories created by the covariates in the model. For instance, if there were two covariates, say gender and smoking status, the covariate patterns would be: male smoker, female smoker, male non-smoker, female non-smoker.

Pearson's test is usually called Pearson's chi-squared test, and the test statistic often represented by X^2. The name describes the distribution of the statistic. Pearson proved that, given a single variate with some fixed number of exclusive classes, the asymptotic distribution of the statistic is chi-squared, with degrees of freedom equal to the number of classes, minus one. In the case when there is more than one variate, then the degrees of freedom are calculated by determining the number of classes for each variate, subtracting one from each, and then taking the product. Pearson only considered the case when all parameters are specified. R.A. Fisher extended X^2 to situations where parameters are unspecified. He proved that, in the case of a single variate, the distribution then is asymptotically chi-squared with degrees of freedom equal to the number of possible response types, minus the number of estimated parameters, minus one.

(Kendall, et al. 1999)

In order for the distributional properties of X^2 to hold, there must be enough data in each cell of the contingency table, or likewise with every particular covariate pattern. If there are too few observations of this type recorded, then there is not enough information to estimate the binomial probability for that particular pattern. In this case, the theory behind the asymptotic distribution of X^2 is not valid. For example, this happens when continuous data are included in a logistic regression model. In such cases, the number of covariate patterns may be as large as the number of observations. Using X^2 in this situation is not appropriate (Kendall, et al. 1999, Hosmer and Lemeshow 2000).

A solution to this problem is to create artificial groups and apply a statistic similar to X^2. That is, to form groups via a method that is based on more than just the natural groups formed by the covariate patterns. Hosmer and Lemeshow (1980) were among the first to offer this type of solution. Their statistic, denoted here as HL, uses a method that is based on ordering and placing into groups the predicted probabilities produced from the model. These are the estimated probabilities that the outcome will occur, given the observed covariate data. The number of groups will be denoted here as G. As a practical matter, they recommended creating ten groups and called the method "deciles-of-risk". In practice though, other numbers of groups can be used. Because these groups are created using the estimated parameters that reference all of the data, which are random, the group boundaries are also random (Moore and Spruill 1975, Kendall, et al. 1999). That is, the "cut-points" of the groups vary. This affects the distribution of their statistic. Building on the work of others (Moore 1971, Moore, et al. 1975, Durst 1979), Hosmer and Lemeshow determined that the approximate distribution of HL is $\chi^2(G-2)$.

Although the Hosmer-Lemeshow test is popular, and widely cited in the literature, there have been some difficulties reported. The value of HL calculated using a particular set of data can vary depending on the group boundaries chosen. Also, if tied values of the estimated probabilities are not placed in the same group, different results can be obtained simply by changing the order of the tied observations (Hosmer, Hosmer, Le Cessie and Lemeshow 1997, Pigeon and Heyse 1999a, Bertolini, Damico, Nardi, Tinazzi and Apolone 2000). In addition, if either all of the estimated probabilities in the first or last groups are very tiny or very large, then

the distributional assumptions may not be valid (Pigeon, et al. 1999a). Further, there are typically multiple covariates in a model, and so observations are located in a multidimensional coordinate system. The range of possible locations that any observation can take is within a multidimensional volume. Because predicted probabilities are calculated as a single number, the covariate vector corresponding to each observation in the covariate space is mapped onto a single dimension, sometimes referred to as the "y-space". This can cause problems (Kuss 2002). Points that were far from each other in the volume of the multidimensional covariate space may now be considered close in the single dimensional y-space. By reducing the multidimensional information to a single dimension, the information about how the observations are related in space will likely change, and some information about their original locations in the covariate space will be lost.

A different solution proposed by Tsiatis (1980) involves a statistic, T, that is a quadratic form with a known asymptotic distribution (Halteman 1980, Tsiatis 1980). Rao (2002) defines a quadratic form in n variables, $x_1, x_2, ..., x_n$, as the homogeneous quadratic function of the variables

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j \tag{1.1}$$

where $\mathbf{x} = \left[x_1, x_2, ..., x_n\right]'$ is an $n \times 1$ column vector and $\mathbf{A}$ is an $n \times n$ symmetric matrix. Pearson's statistic is a special case of the Tsiatis quadratic form.

Tsiatis approaches grouping differently than Hosmer and Lemeshow. His solution was to partition the covariate space, thus retaining the original information about which observations are "close" to one another. But there are also problems with this artificial grouping method. There is no free lunch. One obvious problem is that there are many ways to partition the covariate space, and choosing which to use is subjective. Different choices can give different results (Su and Wei 1991). Here is another problem: consider a model containing many covariates, say five. If the partition is performed so that each of these covariates is broken into two groups, then there is a loss of a large amount of information about each covariate. However,

if the partitioning is any finer, the number of groups would increase rapidly. There are already a relatively large number of groups, $2^5 = 32$, created by the coarse partitioning. In order for these groups to be populated sufficiently, for example with five observations each, at least 160 observations must exist. But that assumes that the distribution of observations among groups is even. Since the data are sampled after partitioning and an even distribution cannot necessarily be assumed, large amounts of data sufficient to populate these groups must be collected. This is not ideal. An interesting side note about the Tsiatis test is that it was discovered simultaneously by Halteman (1980), but Tsiatis published first. In his PhD thesis, Halteman gave a more detailed analysis of the problem. He proved that T has a distribution that is $\chi^2(G-1)$, and that its distribution is unaffected when the group boundaries are created with a method that references the data. He performed simulations to verify that the distribution of T was still approximately $\chi^2(G-1)$ for finite data samples when the deciles-of-risk grouping method was used. He found that even though the grouping method referenced the data, the distribution was essentially the same. His work, however, was not published.

Another statistic, one that combines properties of both HL and T, is one that was proposed by Pigeon and Heyse (1999b), and which they refer to as J^2. This statistic is similar to HL, but multiplied by a "correction factor". This results in a common numerator but differing denominators between the two statistics. By multiplying by this factor, Pigeon and Heyse account for variations between the predicted probabilities within the groups. They state that many grouping methods can be used with their statistic, presumably without changing its approximate $\chi^2(G-1)$ distribution. They do not discuss how using grouping methods that reference the data, such as the deciles-of-risk, might affect the distribution of J^2, though they apparently apply this method to J^2 in an example where they compare HL and J^2.

When a new statistic is proposed, it is important to perform simulation studies to assess its calibration and performance. Specifically, it is important to investigate whether the statistic maintains the expected Type I error rate when a correctly specified model is applied to data, and whether it has sufficient power to detect an incorrectly specified model. When it comes to

goodness-of-fit tests for logistic regression with continuous covariates, HL can be considered to be a kind of industry standard (Kuss 2002). When a new goodness-of-fit statistic for logistic regression with continuous covariates is proposed in the literature, HL is often included in simulation studies to provide a comparison to the new statistic. However, to our knowledge, no simulations studies have been published in the literature that compare HL, J^2, and T. This leads to the first set of research questions addressed in this thesis.

1.2 Research Questions

The first set of research questions that are addressed in this thesis are:

> *If the deciles-of-risk grouping method is applied to J^2 and to T, are their reported distributions unaffected?*
>
> *If the same grouping method (deciles-of-risk) is applied to HL, J^2, and T, are there any differences in their performances? Specifically, do they all maintain the expected Type I error rate, and do they have similar power to detect incorrectly specified models?*

A second set of research questions addressed in this thesis regards the application of HL, J^2, and T to GLMs with binary outcomes other than the logistic regression model. The logistic model is the canonical GLM when the outcomes are assumed to come from a Bernoulli distribution (McCullagh, et al. 1989). There are other models that can also be fit to Bernoulli data, including the probit, the log-log, the complementary log-log, and the log binomial models. These are all non-canonical GLMs (McCullagh, et al. 1989, Hardin and Hilbe 2007). These will be described in more detail in Chapter 2. If any of these models are chosen to relate the outcome and the covariates, the question still remains, "Does this model fit the data well?" Few goodness-of-fit test statistics have been studied for these non-canonical GLMs. Some work has been done, including a study by Blizzard and Hosmer (2006). They applied HL to the binary log binomial model. They wondered whether the distributional assumptions that apply when assessing the logistic model still apply in the log binomial case. Although they found evidence

that it could be applied, they concluded that more research was needed. This leads to the other research questions addressed in this thesis. They are

Can HL, J^2, and T be used as goodness-of-fit test statistics for non-canonical GLMs with Bernoulli outcomes? If so, what are their distributions under these models?

If HL, J^2, and T are used as goodness-of-fit statistics for non-canonical GLMs with Bernoulli outcomes and continuous covariates, are there differences in their performances?

1.3 Organization of Thesis

This thesis is organized as follows: In Chapter two, necessary notation is introduced and background concepts are discussed. Chapter three contains a review of the literature on goodness-of-fit statistics for Bernoulli GLMs. Chapter four includes an analytical treatment and a simulation study comparing the HL, J^2, and T goodness-of-fit statistics in the logistic model setting when the deciles-of-risk grouping method is applied. An example comparing the statistics in the logistic model setting is also presented. A new goodness-of-fit statistic for all GLM with Bernoulli outcomes, based on T, is introduced in Chapter five, along with corresponding forms of HL and T. In addition, a simulation study comparing the distributional characteristics and performance of the new statistic under the probit, log-log, complementary log-log, and log binomial models is performed. Chapter five concludes with a real-world example that illustrates the use of the three statistics under the four non-canonical link functions studied. Finally, Chapter six contains an overall discussion of the research, including its importance and limitations, as well as suggestions for future research.

Chapter 2 Notation and Basic Concepts

2.1 Notation

First, we will set some notation to be used in the following chapters. Let $\mathbf{Y} = (Y_1,...,Y_n)'$ represent a column vector of n outcome random variables, and $\mathbf{y} = (y_1,...,y_n)'$ represent the column vector of observed outcomes, which are realizations of $\mathbf{Y}$. Assume further that the mean values of the components of $\mathbf{Y}$, $\boldsymbol{\mu} = (\mu_1,...,\mu_n)'$, can be modelled with a function that depends on the values of K covariates and a constant. Consider n independent observations of the pairs $(\mathbf{x}_i, y_i)$, $i = 1,...,n$, where, for the ith observation, y_i is the observed outcome of the random variable Y_i, and $\mathbf{x}_i = (x_{i0}, x_{i1},..., x_{iK})'$ is a column vector of the observed covariate values, plus $x_{i0} = 1$ to allow for the estimation of a regression constant. Let an $n \times (K+1)$ matrix $\mathbf{X}$ represent a design matrix, where each row contains the covariate data, $\mathbf{x}_i'$, for each of the n observations. Finally, let $\boldsymbol{\beta} = (\beta_0, \beta_1,..., \beta_K)'$ represent a column vector of regression coefficients. Other notation will be introduced within the text.

2.2 Generalized Linear Models

During the 20th century, several regression models were developed for the analysis of a variety of data types, each requiring a different maximum likelihood algorithm for the estimation of model coefficients and standard errors (Hardin, et al. 2007). Examples include the Gaussian or normal model, the Poisson model, the probit model, and the logit model. The methods required to estimate the parameters of some of these models are mathematically intensive. However, the methods required to estimate the parameters of the Gaussian linear model, the Ordinary Least Squares (OLS) method which has less restrictive assumptions than the maximum likelihood methods, are relatively straightforward. Nelder and Wedderburn (1972) introduced the GLM that unifies several of these well-known models into a single class. Through the GLM theory,

the relatively simple OLS methods for the estimation of model coefficients are extended to all members of this class. The basic strategy of the GLM is to choose a link function that relates $\boldsymbol{\mu}$ to a set of linear predictors, $\boldsymbol{\eta} = (\eta_1,...,\eta_n) = \mathbf{X}\boldsymbol{\beta}$ (McCullagh, et al. 1989). The following are assumed under GLM theory:

1) the components of $\mathbf{Y}$ are independently distributed, have means $\boldsymbol{\mu} = (\mu_1,...,\mu_n)$, and each has a distribution in the exponential family;
2) there is a systematic component, $\boldsymbol{\eta} = (\eta_1,...,\eta_n)' = \mathbf{X}\boldsymbol{\beta}$, which is a column vector of linear predictors, with the ith element expressed as $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$; and
3) there is a monotonic, differentiable link function, $g(\cdot)$, that relates $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$, such that, for the ith observation, $g(\mu_i) = \eta_i$

(McCullagh, et al. 1989).

2.3 Exponential Family

The outcomes modelled by GLMs are assumed to have distributions in the exponential family. Some examples of exponential family distributions are the Gaussian, binomial, and Poisson. For a distribution to be an exponential family member, it must be possible to express the probability function of the outcome random variable, Y, in the form

$$f_Y(y;\zeta,\phi) = \exp\left\{\frac{y\zeta - b(\zeta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

where ζ is the canonical parameter, $b(\zeta)$ is the cumulant function, ϕ is the dispersion parameter, and $c(y,\phi)$ is a normalizing term (McCullagh, et al. 1989, Hardin, et al. 2007). Note that by definition the terms $y_i\zeta_i$ and $b(\zeta_i)$ are both functions of ζ, but not functions of ϕ, while $c(y,\phi)$ and $a(\phi)$ are functions of ϕ, but not of ζ. The normalizing term, $c(y,\phi)$, is only used to scale $f(y \mid \zeta,\phi)$ so that it integrates to 1. The function $a(\phi)$ is used to produce standard errors for some of the exponential family distributions. When the chosen link function

of the GLM is the same function as the canonical parameter ζ, then $\zeta = \eta$, and the link function is referred to as the canonical link.

The exponential family of distributions have been historically popular partly because of their many useful algebraic properties. One result of these properties is that the first derivative of the log-likelihood function of each exponential family distributions results in an "observed minus expected" form. Goodness-of-fit test statistics for regression models typically are based on the comparison of the observed outcomes and the expected outcomes produced by the model, and so GOF statistics can be developed that are based on the derivatives of the log-likelihood functions of the exponential family distributions. We describe the general form of the likelihood function for GLM here, as well as some identities that will be used later when discussing GOF statistics.

The joint density function of an exponential family distribution for a set of outcomes $\mathbf{y}$, given the canonical parameter and dispersion parameter, is expressed as

$$f(\mathbf{y} \mid \zeta, \phi) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i \zeta_i - b(\zeta_i)}{a(\phi)} + c(y_i, \phi) \right\} \tag{2.2}$$

since the observations are considered to be independent. The likelihood function is equal to the probability density function, and has the same form as (2.2), but differs in that it conditions on the observed data rather than on the parameters. That is, it indicates how likely it would be to observe the sampled outcomes as a function of possible values of the canonical parameter, ζ, and ϕ. For n independent and identically distributed (i.i.d.) observations, the likelihood function is expressed as

$$L = \prod_{i=1}^{n} f(\zeta, \phi \mid y_i) \tag{2.3}$$

The log of the likelihood function, l, is often easier to work with mathematically. Assuming that the observations are i.i.d., the joint log likelihood for members of the exponential family is expressed as

$$l = \sum_{i=1}^{n} \left\{ \frac{y_i \zeta_i - b(\zeta_i)}{a(\phi)} + c(y_i, \phi) \right\} \tag{2.4}$$

Next we present some of the identities that will be used later in the thesis. Their derivations can be found in Kendall, et al. (1999) (section 17.14) and in Hardin, et al. (2007).

The mean or expected value of the first partial derivative of the log likelihood is

$$E\left(\frac{\partial l}{\partial \zeta} \right) = 0 \tag{2.5}$$

and

$$E\left(\frac{\partial^2 l}{\partial \zeta^2} + \left(\frac{\partial l}{\partial \zeta} \right)^2 \right) = 0 \tag{2.6}$$

Assuming that the likelihood function is twice differentiable, differentiating both sides of (2.5) and using (2.6), gives

$$\begin{aligned} \text{var}\left(\frac{\partial l}{\partial \zeta} \right) &= E\left\{ \left(\frac{\partial l}{\partial \zeta} \right)^2 - \frac{\partial^2 l}{\partial \zeta^2} \right\} \\ &= -E\left(\frac{\partial^2 l}{\partial \zeta^2} \right) \end{aligned} \tag{2.7}$$

and

$$E\left\{ \left(\frac{\partial l}{\partial \zeta} \right)^2 \right\} = -E\left(\frac{\partial^2 l}{\partial \zeta^2} \right) \tag{2.8}$$

The left hand side of (2.8) is sometimes called the Fisher Information ((Kendall, et al. 1999) Section 17.15).

Using (2.4), it can be shown (McCullagh, et al. 1989) that

$$\frac{\partial b(\zeta_i)}{\partial \zeta} = E(y_i) = \mu_i \tag{2.9}$$

Further, using (2.6) , and (2.9), it can be shown (McCullagh, et al. 1989) that

$$\frac{\partial^2 b(\zeta_i)}{\partial\zeta\partial\zeta} = \frac{\text{var}(y)}{a(\phi)} \tag{2.10}$$

McCullagh, et al. (1989) note that because the term $\partial^2 b(\zeta_i)/\partial\zeta\partial\zeta$ depends only on the canonical parameter, and thus the mean, it is sometimes called the variance function and they express (2.10) as $var(\mu)$.

The chain rule can be applied to $\partial l/\partial\zeta$ to obtain the resulting score vector, $(\partial l/\partial\beta_0, \partial l/\partial\beta_1, ..., \partial l/\partial\beta_K)'$. Hardin, et al. (2007) show in detail that, for the kth regression coefficient, the score is

$$\frac{\partial l}{\partial\beta_k} = \sum_{i=1}^{n}\left(\frac{y_i - \mu_i}{a(\phi)\text{var}(\mu_i)}\right)\left(\frac{\partial\mu_i}{\partial\eta_i}\right)x_{ik} \tag{2.11}$$

They also show that if the link function chosen is the canonical link function, where $\zeta = \eta$, then (2.11) simplifies to

$$\frac{\partial l}{\partial\beta_k} = \sum_{i=1}^{n}\left(\frac{y_i - \mu_i}{a(\phi)}\right)x_{ik} \tag{2.12}$$

Setting (2.11) to zero for each of the scores, results in $K+1$ estimating equations. The roots of the $K+1$ estimating equations can be used to estimate the values that maximize the log likelihood function, and thus determine the maxima of the likelihood (Kendall, et al. 1999). The maximum likelihood estimate of β_k are expressed here as $\hat{\beta}_k$.

There are two main methods used to estimate the regression coefficients of GLM. We give a brief overview here, but a more detailed discussion can be found in Hardin, et al. (2007). The first is an iterative numeric search method called Newton-Raphson. When estimating the regression coefficients, this method uses the Observed Information Matrix (OIM). The OIM is the negative of the observed Hessian matrix. For the exponential family, the OIM has elements expressed generally as

$$-\frac{\partial^2 l}{\partial\beta_k\partial\beta_{k'}} = \sum_{i=1}^{n}\frac{1}{a(\phi)}\left[\frac{1}{\left(\mathrm{var}(\mu_i)\right)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 + (y_i-\mu_i)\left(\frac{1}{\left(\mathrm{var}(\mu_i)\right)^2}\frac{\partial\,\mathrm{var}(\mu_i)}{\partial\mu_i}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 - \left(\frac{1}{\mathrm{var}(\mu_i)}\right)\left(\frac{\partial^2\mu_i}{\partial\eta_i^2}\right)\right)\right]\frac{\partial\eta}{\partial\beta_k}\frac{\partial\eta}{\partial\beta_{k'}} \tag{2.13}$$

The covariance matrix of the model parameters is the inverse of the OIM. (Searle 2006, Hardin, et al. 2007).

A second iterative numeric method for estimating the regression coefficients, due to Fisher, is called "the method of scoring" or the iteratively reweighted least squares (IRLS) method (Kendall, et al. 1999, Hardin, et al. 2007). In this case, instead of the OIM, the IRLS method uses the expected information matrix (EIM), or Fisher information in (2.8). The EIM has elements expressed generally as

$$E\left(-\frac{\partial^2 l}{\partial\beta_k\partial\beta_{k'}}\right) = \sum_{i=1}^{n}\frac{1}{a(\phi)}\left[\frac{1}{\mathrm{var}(\mu_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\right]\frac{\partial\eta}{\partial\beta_k}\frac{\partial\eta}{\partial\beta_{k'}} \tag{2.14}$$

When the link of the GLM is the canonical link function, that is when $\zeta = \eta$, the matrices (2.13) and (2.14) are equal, but this is not the case under non-canonical GLM.

Some test statistics, such as Rao's efficient score test (Rao 2002), are based on the identities (2.5) and (2.7). Goodness-of-fit statistics for canonical GLM can be developed from score tests when the selected link function for the GLM is canonical, since then the estimating equations reduce to the form (2.12). (McCullagh, et al. 1989, Kendall, et al. 1999). This is not the case when the link function of the GLM is non-canonical.

2.4 GLM for Binary Data

Some of the most commonly used GLMs are those whose random components have binomial distributions (Hardin, et al. 2007). We are primarily concerned with GLMs for binary outcome data, and so first present here the properties of the binomial likelihood function that were discussed in sections 2.2 and 2.3, and then specify the same for the Bernoulli case.

The probability mass function of the binomial distribution can be expressed as

$$f\left(y \mid n, p\right) = \binom{n}{p} p^{y} \left(1-p\right)^{n-y} \tag{2.15}$$

where p is the probability of a successful outcome of an experiment and n is the number of times the experiment was performed. Because the binomial distribution is an exponential family member, its distribution can be rewritten in the form of (2.1) as

$$f\left(y \mid n, p\right) = \exp\left\{ y \ln\left(\frac{p}{1-p}\right) + n \ln\left(1-p\right) + \ln\binom{n}{p} \right\} \tag{2.16}$$

In the case of the binomial distribution, the canonical parameter is

$$\zeta = \ln\left(\frac{p}{1-p}\right) \tag{2.17}$$

the cumulant function is $b\left(\zeta\right) = -n \ln\left(1-p\right)$, the dispersion parameter is $\phi = 1$, and the normalizing term is $c\left(y, \zeta\right) = 0$ (McCullagh, et al. 1989). A special case of the binomial distribution, when $n = 1$, is the Bernoulli distribution.

The link function chosen, which relates a binomial random variable Y to a linear predictor η, may be any suitable function that is monotonic and differentiable (McCullagh, et al. 1989). However, if the canonical parameter of the exponential family is chosen as the link function it leads to several desirable statistical properties, including relatively straightforward methods for parameter estimation and methods for evaluating model fit. When Y comes from a Bernoulli distribution the canonical link function is the binary logit function,

$$g\left(\mu\right) = \ln\left(\frac{p}{1-p}\right) \tag{2.18}$$

(McCullagh, et al. 1989). In this case, as with all canonical link functions, $g\left(\mu\right) = \eta$.

The mean and variance of y for GLM with Bernoulli outcomes can be obtained regardless of link function used, (Hardin, et al. 2007) by taking the first and second derivatives of the cumulant function,

$$\frac{\partial b}{\partial \zeta} = p \tag{2.19}$$

and

$$\frac{\partial^2 b}{\partial \zeta \partial \zeta} = p(1-p) \tag{2.20}$$

Thus, as noted by Hardin, et al. (2007), unlike the linear model, when the random outcomes are binomially distributed, the variance is related to the mean and thus is not constant, but varies with the mean.

The Bernoulli probability density function expressed in canonical exponential form is

$$f(y \mid \zeta, \phi) = \exp\left\{ y \ln\left(\frac{p}{1-p} \right) + \ln(1-p) \right\} \tag{2.21}$$

and the joint log likelihood in canonical form, is

$$l(p; y) = \sum_{i=1}^{n} y_i \ln\left(\frac{p_i}{1-p_i} \right) + \ln(1-p_i) \tag{2.22}$$

By (2.11), (2.19), and (2.20), the *k*th element of the score vector is

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{n} \left(\frac{y_i - p_i}{p_i(1-p_i)} \right) \left(\frac{\partial p_i}{\partial \eta_i} \right) x_{ki} \tag{2.23}$$

For Bernoulli outcomes, the general terms of the OIM, (2.13), are

$$\begin{aligned} -\frac{\partial^2 l}{\partial \beta_k \partial \beta_{k'}} = \sum_{i=1}^{n} \Bigg[& \frac{1}{p_i(1-p_i)} \left(\frac{\partial p_i}{\partial \eta_i} \right)^2 \\ & + (y_i - p_i) \left(\frac{1}{\left(p_i(1-p_i) \right)^2} (1-2p_i) \left(\frac{\partial p_i}{\partial \eta_i} \right)^2 - \left(\frac{1}{p_i(1-p_i)} \right) \left(\frac{\partial^2 p_i}{\partial \eta_i^2} \right) \right) \Bigg] \frac{\partial \eta}{\partial \beta_k} \frac{\partial \eta}{\partial \beta_{k'}} \end{aligned} \tag{2.24}$$

and the terms of the EIM, (2.14), are

$$-E\left(\frac{\partial^2 l}{\partial \beta_k \partial \beta_{k'}} \right) = \sum_{i=1}^{n} \left[\frac{1}{p_i(1-p_i)} \left(\frac{\partial p_i}{\partial \eta} \right)_i^2 \right] \frac{\partial \eta}{\partial \beta_k} \frac{\partial \eta}{\partial \beta_{k'}} \tag{2.25}$$

2.5 Canonical GLM for Bernoulli Outcomes (Binary Logistic Regression)

Generalized linear models with Bernoulli outcomes and a canonical logit link function are referred to as logit or logistic models. These are arguably the most commonly used models for relating a binary outcome to a set of predictor variables (Hardin, et al. 2007). This model is central to the discussion in Chapter 4 and Chapter 5, and we present the expressions necessary for our discussion here.

Denoting the $E(Y|\mathbf{x})$ for the logistic model as $\pi(\mathbf{x})$, the canonical logit link function is then

$$g(\pi(\mathbf{x})) = \ln\left\{\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right\} = \eta \tag{2.26}$$

The inverse of the logit link function produces the mean, and is expressed as

$$g^{-1}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)} = \pi(\mathbf{x}) \tag{2.27}$$

The canonical parameter, when the canonical logit link is chosen, is

$$\zeta = \ln\left\{\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right\} = \eta \tag{2.28}$$

and its derivative with respect to η is

$$\frac{\partial\zeta}{\partial\eta} = \frac{\partial\eta}{\partial\eta} = 1 \tag{2.29}$$

In the univariate case, when $\pi(x)$ is graphed as a function of x, a sigmoidal curve is produced that is symmetric about 0.5. An example is shown in Figure 2.1. In the binary case, this property allows for the code designation of one of the outcomes as a success and the other as a failure to be reversed (with only a change of sign).

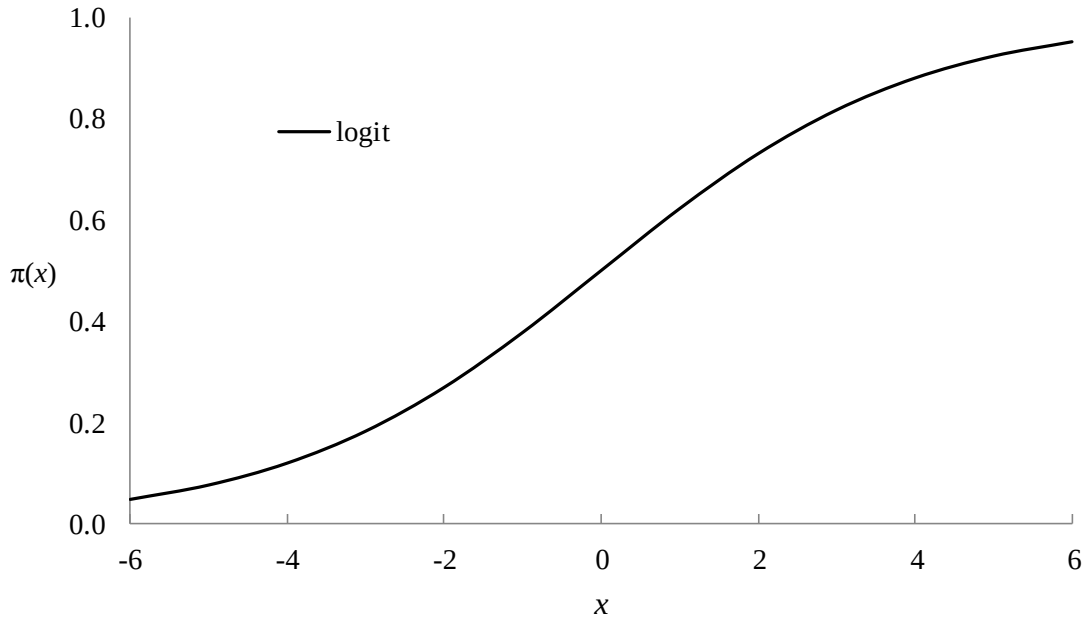

Figure 2.1 Graph of the probabilities produced by the inverse of the logit link function as a function of x, where $x \sim U(-6,6)$ and $\eta = 0.5x$.

2.6 Non-Canonical Link Functions for Binary Outcomes

Generalized linear models with non-canonical link functions can also be used to relate outcomes with Bernoulli distributions to a set of explanatory variables. We introduce several of these models here, and the expressions necessary for the development of a goodness-of-fit test statistic for non-canonical binary GLM presented in Chapter 5. In the previous section, the $E(Y \mid \mathbf{x})$ under the canonical logistic model was expressed as $\pi(\mathbf{x})$. When the link function is non-canonical, we will instead designate $E(Y \mid \mathbf{x})$ generally as $\theta_{\mathcal{G}}(\mathbf{x})$. When referring to the specific non-canonical probit, log-log, complementary log-log, and log models, we will replace the subscript and denote the probabilities as $\theta_{\Pr}(\mathbf{x})$, $\theta_{LL}(\mathbf{x})$, $\theta_{Cll}(\mathbf{x})$, and $\theta_{LB}(\mathbf{x})$ respectively.

2.6.1 Probit

The probit model is a GLM with the non-canonical link function

$$g\left(\theta_{\Pr}(\mathbf{x})\right)_{\Pr} = \Phi^{-1}\left(\theta_{\Pr}(\mathbf{x})\right) = \eta \tag{2.30}$$

where Φ is the cumulative normal distribution, expressed as

$$\Phi(\eta) = \int_{-\infty}^{\eta} \phi(v)dv = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\eta} \exp\left(-\frac{1}{2}v^2\right)dv \tag{2.31}$$

Here, $\pi = 3.1415...$, and $\phi(\cdot)$ is the probability density function of the standard normal distribution (Hardin, et al. 2007).

The inverse of the link function giving $E(Y|\mathbf{x})$ is

$$g^{-1}(\eta)_{\text{Pr}} = \Phi(\eta) = \theta_{\text{Pr}}(\mathbf{x}) \tag{2.32}$$

The canonical parameter under the probit model is

$$\zeta = \ln\left(\theta_{\text{Pr}}(\mathbf{x})/1-\theta_{\text{Pr}}(\mathbf{x})\right) = \ln\left(\Phi(\eta)/1-\Phi(\eta)\right) \tag{2.33}$$

and its derivative with respect to η is

$$\frac{\partial\zeta}{\partial\eta} = \frac{\partial \ln\left[\Phi(\eta)/\{1-\Phi(\eta)\}\right]}{\partial\eta} \tag{2.34}$$

Like the logistic model, the probit model is symmetrical around a mean probability of 0.5, and has a similar sigmoidal shape. An example when the inverse of the probit link for a univariate model, $E(Y|x)$, is plotted against x is shown in Figure 2.2.

2.6.2 Log-log and Complementary Log-log

If the binary data to be analysed are unbalanced, that is, the number of 0s and 1s are unequal, then either the log-log model or the complementary log-log model may be a better fit. If there are many more 0s than 1s, then the log-log model may be the more appropriate model, whereas if there are many more 1s than 0s, then the complementary log-log model may give a better fit to the data. The $E(Y|\mathbf{x})$ for these models is denoted here as $\theta_{LL}(\mathbf{x})$ and $\theta_{Cll}(\mathbf{x})$ respectively. The link functions for the log-log and complementary log-log models are

$$g(\mu)_{LL} = -\ln\left\{-\ln\left(\theta_{LL}(\mathbf{x})\right)\right\} = \eta \tag{2.35}$$

and

$$g(\mu)_{Cll} = \ln\{-\ln(1-\theta_{Cll}(\mathbf{x}))\} = \eta \tag{2.36}$$

respectively. The corresponding inverses of these link functions are

$$g^{-1}(\eta)_{LL} = \exp\{-\exp(-\eta)\} = \mu = \theta_{LL}(\mathbf{x}) \tag{2.37}$$

and

$$g^{-1}(\eta)_{Cll} = 1-\exp\{-\exp(\eta)\} = \mu = \theta_{Cll}(\mathbf{x}) \tag{2.38}$$

The canonical parameters under the log-log and complementary log-log models are

$$\zeta = \ln\left(\frac{\theta_{LL}(\mathbf{x})}{1-\theta_{LL}(\mathbf{x})}\right) \tag{2.39}$$

and

$$\zeta = \ln\left(\frac{\theta_{Cll}(\mathbf{x})}{1-\theta_{Cll}(\mathbf{x})}\right) \tag{2.40}$$

respectively. Their derivatives with respect to η are

$$\frac{\partial\zeta}{\partial\eta} = \frac{\partial\ln\left[\theta_{LL}(\mathbf{x})/\{1-\theta_{LL}(\mathbf{x})\}\right]}{\partial\eta}$$

$$= -\frac{\ln\{\theta_{LL}(\mathbf{x})\}}{1-\theta_{LL}(\mathbf{x})} \tag{2.41}$$

and

$$\frac{\partial\zeta}{\partial\eta} = \frac{\partial\ln\left[\theta_{Cll}(\mathbf{x})/\{1-\theta_{Cll}(\mathbf{x})\}\right]}{\partial\eta}$$

$$= -\frac{\ln\{1-\theta_{Cll}(\mathbf{x})\}}{\theta_{Cll}(\mathbf{x})} \tag{2.42}$$

respectively.

These models produce asymmetric sigmoidal curves. The log-log curve has an elongated lower tail, and the complementary log-log has an elongated upper tail. Examples of both the graph of

the inverse of the log-log and the complementary log-log links for a univariate model, $E(Y \mid x)$, as a function of x are shown in Figure 2.2.

Because of the inherent asymmetry of these models, the coding of the model is not symmetric. That is, the assignment of "success" and "failure" cannot be simply reversed (with just a sign change). However the log-log and the complementary log-log make a complementary pair. That is, the log-log model applied to outcomes y will give the same result as the complementary log-log model applied to $1-y$. Similarly, if the log-log is applied to $1-y$, then the same result is obtained when the complementary log-log is applied to y.

2.6.3 Log Binomial

Another non-canonical GLM for binary data is the log model, also known as the log binomial model. The $E(Y \mid \mathbf{x})$ for the log binomial model will be denoted here as $\theta_{LB}(\mathbf{x})$. The link function for the log binomial model is the natural log function, and is expressed as,

$$g\left(\theta_{LB}(\mathbf{x})\right)_{LB} = \ln\left(\theta_{LB}(\mathbf{x})\right) = \eta \tag{2.43}$$

The inverse of the link function is

$$g^{-1}(\eta)_{LB} = \exp(\eta) = \theta_{LB}(\mathbf{x}) \tag{2.44}$$

The canonical parameter under the log binomial model is

$$\zeta = \ln\left[\theta_{LB}(\mathbf{x}) / \left\{1-\theta_{LB}(\mathbf{x})\right\}\right] \tag{2.45}$$

and its derivative with respect to η is

$$\frac{\partial \zeta}{\partial \eta} = \frac{\partial \ln\left[\theta_{LB}(\mathbf{x}) / \left\{1-\theta_{LB}(\mathbf{x})\right\}\right]}{\partial \eta}$$

$$= \frac{1}{1-\theta_{LB}(\mathbf{x})} \tag{2.46}$$

The lower tail of the inverse log link function when plotted against x is similar to that of the logit, but the upper tail can be very different since the resulting values of $E(Y \mid x)$ may be

greater than 1. An example of the graph of the inverse of the log link for a univariate model as a function of x is shown in Figure 2.2.

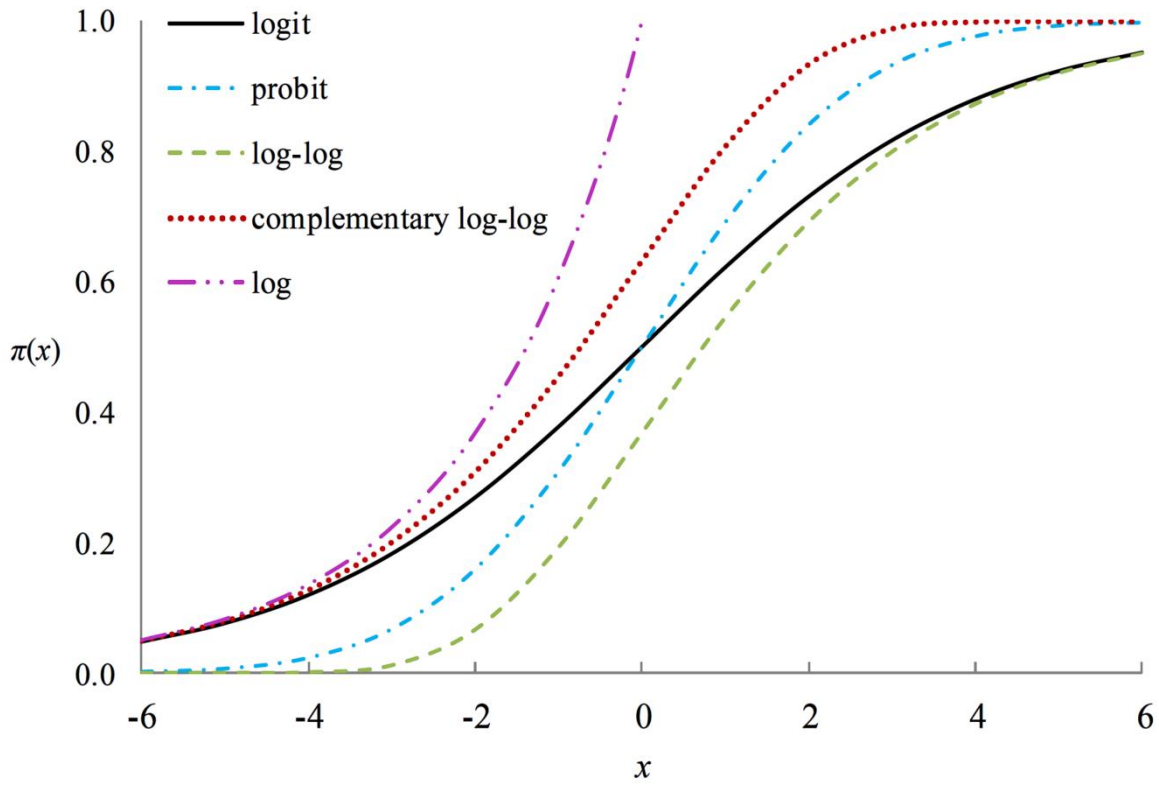

Figure 2.2 A graph of the probabilities produced by the inverse of five link functions as a function of x. ($x \sim U(-6,6)$ and $\eta = 0.5x$)

2.7 Basic Concepts of Score Tests

Some goodness-of-fit tests are formed as score tests. The score test statistic was introduced by Rao (1948), and is based on the identities (2.5) and (2.6) (Smyth 2003). Suppose that the likelihood function of a GLM, (2.4), depends on two vectors of parameters, and express this as $l(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y})$. The null hypothesis being tested is that $\boldsymbol{\gamma} = \mathbf{0}$, against the alternative hypothesis that $\boldsymbol{\gamma}$ is unrestricted. Here, the $\boldsymbol{\beta}$ are considered nuisance parameters, while the $\boldsymbol{\gamma}$ are considered the parameters of interest. The column vector of scores can be partitioned as

$$\mathbf{S} = \left(\mathbf{S}_{\beta}, \mathbf{S}_{\gamma}\right)' \tag{2.46}$$

with

$$\mathbf{S}_{\beta} = \frac{\partial l}{\partial \boldsymbol{\beta}} \tag{2.46}$$

and

$$\mathbf{S}_{\gamma} = \frac{\partial l}{\partial \boldsymbol{\gamma}} \tag{2.46}$$

The covariance matrix of the score vector (2.46) is the expected information matrix, $\mathbf{I}$, with general elements as in (2.14). It can be partitioned to conform with (2.46) and (2.46), and expressed as

$$\mathbf{I} = \begin{bmatrix} \mathbf{I}_{\beta\beta} & \mathbf{I}_{\beta\gamma} \\ \mathbf{I}_{\gamma\beta} & \mathbf{I}_{\gamma\gamma} \end{bmatrix} \tag{2.47}$$

(Graybill 1976, McCullagh, et al. 1989). If the values of the nuisance parameters, $\boldsymbol{\beta}$, are known, then the score test statistic, evaluated under the null hypothesis, is

$$\mathbf{S}_{\gamma}' \left(\mathbf{I}_{\gamma\gamma} \right)^{-1} \mathbf{S}_{\gamma} \Big|_{\boldsymbol{\gamma}=\mathbf{0}} \tag{2.47}$$

(Smyth 2003). In this case, the score test statistic has a chi-squared distribution with degrees of freedom equal to the rank of $\mathbf{I}_{\gamma\gamma}$ (Rao 2002).

If the nuisance parameters are not known, then the maximum likelihood estimates under the null hypothesis, $\hat{\boldsymbol{\beta}}$, are substituted. Using the multivariate theory on inverses (Graybill 1976, McCullagh, et al. 1989), the covariance matrix, conditional on $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_0$, is

$$\mathbf{I}_{\gamma|\hat{\beta}} = \mathbf{I}_{\gamma\gamma} - \mathbf{I}_{\gamma\beta} \mathbf{I}_{\beta\beta}^{-1} \mathbf{I}_{\beta\gamma} \tag{2.48}$$

(McCullagh, et al. 1989, Smyth 2003).

The score test statistic in this case is calculated as

$$\mathbf{S}_{\gamma}' \left(\mathbf{I}_{\gamma|\hat{\beta}} \right)^{-1} \mathbf{S}_{\gamma} \Big|_{\boldsymbol{\gamma}=\mathbf{0},\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \tag{2.48}$$

where all of the terms of the statistic are evaluated under the null hypothesis $\boldsymbol{\gamma} = \mathbf{0}$, and under $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. Here, the score test statistic has a distribution that is asymptotically chi-squared with degrees of freedom equal to the rank of $\mathbf{I}_{\gamma|\hat{\beta}}$ (Rao 2002) . If $\mathbf{I}_{\beta\gamma} = \mathbf{0}$, then $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are orthogonal, and thus $\mathbf{S}_{\beta}$ and $\mathbf{S}_{\gamma}$ are independent, and the statistics (2.47) and (2.48) are equal (Smyth 2003).

Chapter 3 Literature Review

3.1 Goodness-of-Fit Statistics for Binary Logistic Regression Models

After selecting the explanatory variables to include in a binary logistic regression model and determining an appropriate mathematical function that best describes an "average" value of the outcome variable for the given covariate values, a critical next step in model development is to examine whether the predicted probabilities calculated from the chosen model and covariates are significantly different from the observed outcome values. That is, does the model fit the observed outcome data well? Goodness-of-fit statistics are used to test the hypothesis that the distribution function of the observed outcome variable is the same as the distribution function of the hypothesized model (Kendall, et al. 1999). Two types of goodness-of-fit tests, specific and global, are used to test this null hypothesis (Kuss 2002). The specific test is used to test the null against the alternative that that another specific model would fit the data better. The global test does not evaluate a specific alternative, but rather tests against the alternative that the model does not fit, without specifying in what way the fit may be poor. The focus of this thesis is on global GOF tests. Two well-known global test statistics used to assess the fit of binary logistic regression models are the deviance and the Pearson's chi-square.

3.1.1 Deviance

The deviance is a likelihood ratio test that compares the fits of two models to the observed data; the first is the null model which is under consideration, and the second is a saturated model in which the null model is nested. The saturated model includes a parameter for each of the J covariate patterns. The hypothesis being tested is that all parameters of the saturated model that are not in the working model are equal to zero. In this section we will refer generally to the estimated probabilities of the binary model as $\hat{\pi}$, for ease of notation. The deviance residual can be expressed as

$$d\left(y_j,\hat{\pi}_j\right)=\pm\left\{2\left[y_j\ln\left(\frac{y_j}{n_j\hat{\pi}_j}\right)+\left(n_j-y_j\right)\ln\left\{\frac{\left(n_j-y_j\right)}{n_j\left(1-\hat{\pi}_j\right)}\right\}\right]\right\}^{1/2} \tag{3.1}$$

where $\hat{\pi}_j$ is the predicted probability for jth covariate pattern, n_j is the number of observations with the jth covariate pattern, and y_j is the number of observations from the n_j subjects with the response $y=1$. The sign of (3.1) is the same as that of $y_j-n_j\hat{\pi}_j$. In cases where $y_j=0$, (3.1) is defined as $d\left(y_j,\hat{\pi}_j\right)=-\sqrt{2n_j\left|\ln\left(1-\hat{\pi}_j\right)\right|}$. When $y_j=n_j$, then it is defined as $d\left(y_j,\hat{\pi}_j\right)=\sqrt{2n_j\left|\ln\left(\hat{\pi}_j\right)\right|}$. When testing the fit of a binary logistic regression model with K fitted covariates and J covariate patterns, the deviance statistic is expressed as

$$D=\sum_{j=1}^{J}d\left(y_j,\hat{\pi}_j\right)_j^2 \tag{3.2}$$

When $n_j\hat{\pi}_j$ is not small, the deviance has an asymptotic distribution that is $\chi^2\left(J-K-1\right)$ (Kendall, et al. 1999, Hosmer, et al. 2000, Agresti 2007).

3.1.2 Pearson's Chi-Squared

Pearson's chi-squared goodness-of-fit statistic, X^2, is a quadratic form with a known asymptotic distribution when the outcome is from a multinomial distribution with mutually exclusive classes, and there are sufficient numbers of observations of both the observed and expected outcomes for each class (Pearson 1900). Using the notation described in section 3.1.1, Pearson's residual can be expressed as

$$r\left(y_j,\hat{\pi}_j\right)=\frac{y_j-n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)}} \tag{3.3}$$

and the Pearson statistic as

$$X^2=\sum_{j=1}^{J}r\left(y_j,\hat{\pi}_j\right)^2 \tag{3.4}$$

That is,

$$X^2 = \sum_{j=1}^{J} \frac{\left\{ y_j - n_j \hat{\pi}_j \right\}^2}{n_j \hat{\pi}_j \left(1 - \hat{\pi}_j\right)} \tag{3.5}$$

(Hosmer, et al. 2000). Like the deviance, when $n_j \hat{\pi}_j$ is not small, X^2 has an asymptotic distribution that is $\chi^2\left(J - K - 1\right)$ (Kendall, et al. 1999, Hosmer, et al. 2000, Agresti 2007).

Originally X^2 was used to test for independence in two-way contingency tables (Pearson 1900, Smyth 2003). Later it was extended for use as a goodness-of-fit test for GLM (McCullagh, et al. 1989, Smyth 2003). Smyth (2003) showed that X^2 is a score test statistic.

3.2 Goodness-of-Fit Statistics for Binary Logistic Regression Models with Continuous Covariates

Both X^2 and D are appropriate for assessing the fit of binary logistic regression models when the covariates are categorical and have a multinomial distribution. However, when the chosen model contains continuous explanatory variables, then the number of classes of the multinomial distribution can increase at near the same rate as the number of observations, and in this case the asymptotic theory behind the distributions of X^2 and D does not hold (Kendall, et al. 1999, Hosmer, et al. 2000). Chernoff and Lehmann (1954) considered the case when the maximum likelihood estimates are based on all n observations rather than on only J classes. They found that under this setting, X^2 has an asymptotic distribution that is between $\chi^2\left(J - K - 1\right)$ and $\chi^2\left(J - 1\right)$ because there is a partial recovery of the K degrees of freedom (Chernoff, et al. 1954, Kendall, et al. 1999).

A further question arises when considering the distribution of X^2 if either the membership in the J classes or the boundaries which define the class cells are determined by referencing the estimated parameters that in turn reference the observations. By doing so, the cells or their boundaries are considered random variables. These additional random variables may affect the distribution of X^2, but the asymptotic theory on which the original X^2 is based does not

account for them (Kendall, et al. 1999). This is of particular concern when the sampled covariate data are continuous, since in this case there is no natural formation of cells. This difficulty was considered first by Watson (1957), and then later by Moore (1971) and Moore, et al. (1975), who presented a unified general theory for chi-squared tests when the cells are random.

3.2.1 Hosmer-Lemeshow Goodness-of-Fit Statistic

Hosmer, et al. (1980) developed a series of goodness-of-fit test statistics for binary logistic regression models when the model evaluated contains continuous covariates. These statistics are all similar in form to X^2, and like X^2, they test the null hypothesis that the model selected fits the data well against the alternative hypothesis that the model does not fit. However, they vary in their assumptions about the distribution of the covariates, their methods of estimating the regression coefficients, and in their strategies for forming groups. Under one set of assumptions, the conditional distribution of $\mathbf{X}$ given Y is assumed to be multivariate normal, while under another set the assumptions are less restrictive. One of two grouping methods for forming group boundary cutpoints was applied when calculating these statistics. The group boundary cutpoints were either fixed or random. The fixed method defines the cutpoints as b/G, where $b = 1, 2, ..., G-1$, and the predicted probabilities are placed into groups if they fall between these cutpoints. Under the random boundary method, often called the "deciles-of-risk" method, the predicted probabilities are ordered and roughly n/G of them are placed into each group. For example, if $G = 10$ and $n = 150$, then about 15 observations would be placed in each group. When tied values occur amongst the predicted probabilities, the ties are placed into the same group, and so it is possible for the numbers of observations within groups to be uneven.

Hosmer and Lemeshow studied four versions of a statistic that uses the fixed cutpoint grouping method. They refer to these statistics as H. The versions of H varied in their assumption about the distribution of the covariates and in the method of regression coefficient estimation. Two other statistics were also studied, and referred to generally as C (with differing subscripts). In this case, grouping was performed using the deciles-of-risk grouping method and the less

restrictive assumptions about the distributions of the covariates were made, but the method of regression coefficient estimation varied. In a later study, Hosmer, Lemeshow and Klar (1988) found that, when the regression coefficient estimates were made using maximum likelihood estimation, C adhered to the assumed distribution better than the other statistics in the original study. This statistic is the one that is cited most widely in the literature, and is sometimes referred to as $\hat{C}$ (Hosmer, et al. 1997, Hosmer, et al. 2000). We refer to it here as HL. Although the grouping method used with HL is the deciles-of-risk method, in fact, other percentiles can be used as long as they not near n. However, Hosmer, et al. (2000) do not recommend calculating HL with less than six groups, because this almost always results in the test indicating that the model fits the data.

The HL statistic is conditional on the maximum likelihood estimates of the parameters, and is expressed as

$$HL = \sum_{g=1}^{G} \frac{\left\{ \sum_{i \in \psi_g} \left(y_i - \hat{\pi}_i \right) \right\}^2}{n_g \bar{\pi}_g \left(1 - \bar{\pi}_g \right)} \tag{3.6}$$

where $i \in \psi_g$ denotes the set of all observations i in the gth group, n_g is the number of observations in the gth group, and $\bar{\pi}_g$ is the mean of the predicted probabilities in the gth group, and is calculated as

$$\bar{\pi}_g = \frac{\sum_{i \in \psi_g} \hat{\pi}_i}{n_g} \tag{3.7}$$

The distribution of HL is affected by two conditions. The first is that the estimates of the regression parameters are determined using likelihood functions for ungrouped data. The second is that the boundaries for any group are dependent on the estimated parameters, and thus the groups are random (Hosmer, et al. 1980). Applying the work of Moore, et al. (1975) and of Durst (1979), Hosmer, et al. (1980) showed that the asymptotic distribution of HL is

$$\chi^2 \left(2G - G - (K+1) \right) + \sum_{k=1}^{K+1} \lambda_k \chi^2{}_k \left(1 \right) \tag{3.8}$$

where K is the number of covariates, and λ_i is the ith non-zero or 1 eigenvalue of the covariance matrix of the HL statistic, $0 < \lambda_k < 1$, $k = 1,...,K$. Through simulations they showed that the contribution of $\sum_{k=1}^{K+1} \lambda_k \chi^2_k(1)$ is approximately $\chi^2(K-1)$, and thus the distribution of HL is approximately $\chi^2(G-2)$.

The Hosmer-Lemeshow test is often used to assess the fit of logistic regression models with continuous covariates, and is widely cited in the literature. This is partly due to its ease of calculation, its intuitive appeal, and its availability in most major statistical packages (e.g. SAS, STATA, and SPSS). It also may be due to the fact that it has been studied extensively. Many simulation studies have been performed to evaluate its performance against other statistics.

Although HL is regularly used when logistic models fit are assessed, some issues have been reported. Two studies (Hosmer, et al. 1997, Pigeon, et al. 1999a) pointed out that different software packages can report different values of HL even though the packages produced the same model. This has been attributed to differences in how the software algorithms define the deciles. Bertolini, et al. (2000) note that when the number of covariate patterns is less than n and there are ties among the predicted probabilities, if the ties are placed into different groups, different values of HL can occur if the order of the observations is varied. Hosmer, et al. (2000), however, point out that the problem of assigning tied values to different groups is usually only an issue when there are few covariate patterns forming the predicted probabilities. In this case, if ties are grouped together there may be some groups that contain many more observations than n / g, or conversely there may be too few in another group.

Pigeon, et al. (1999a) also point out that if the deciles-of-risk partitioning method is used, and the predicted probabilities in a group are either all near 0 or all near 1, the expected frequency of an event (i.e. the sum of the predicted probabilities in that group) or of a non-event may be less than 1. This would invalidate the chi-square approximation for the distribution of HL. One solution is to combine some groups, which will raise the expected frequencies within these groups. This strategy is possible because grouping is performed after the data are collected.

Another criticism of HL is its lack of power to detect certain types of poor fit in a model. Observations grouped together under the deciles-of-risk method can have very different covariate patterns. This happens when observations in multi-dimensional covariate space are mapped onto the single dimensional "y-space". Observations that in some sense were originally close in the covariate space may now be far apart in the "y-space". Because the unique covariate patterns are now represented only through their fitted probabilities, and different covariate patterns may have the same fitted probability, it may be difficult to identify which types of subjects are not represented well by the model.

To overcome some of the deficiencies of the deciles-of-risk grouping strategy, Hosmer, et al. (2000) recommend that HL, which is a summary statistic, be used in conjunction with diagnostic statistics and other methods of fit analysis that evaluate individual residuals. These might include classification tables, ROC curves, R^2, and regression diagnostics for logistic regression based on work in relation to leverage, and diagnostic plots by Pregibon (1981).

Although these deficiencies have been reported, no other methods have been put forward that do not also have difficulties. In consequence, the Hosmer-Lemeshow GOF statistic remains one of the standard goodness-of-fit tests when evaluating the fit of a logistic regression model with continuous covariates.

3.2.2 Tsiatis Goodness-of-Fit Score Statistic

About the same time that Hosmer, et al. (1980) published their work, Tsiatis (1980) introduced another goodness-of-fit statistic that can be used to evaluate the fit of a binary logistic regression model with continuous covariates. It is a score test with nuisance parameters, as described in section 2.7. First, the covariate space is partitioned into G distinct regions without reference to the estimated parameters or observed data, and thus the partitions are not random. Then an augmented logistic model is introduced that gives the conditional probability of a successful outcome, given the observed values of the covariates, as

$$\boldsymbol{\pi}(\mathbf{x},\mathbf{I}) = \frac{\exp\left(\mathbf{x}'\boldsymbol{\beta} + \sum_{g=1}^{G}\gamma_g I^{(g)}\right)}{1+\exp\left(\mathbf{x}'\boldsymbol{\beta} + \sum_{g=1}^{G}\gamma_g I^{(g)}\right)} \tag{3.9}$$

where $\left\{I^{(1)},...,I^{(G)}\right\}$ are a set of indicator functions that are defined as $I^{(g)}=1$ when the covariates lie in it the gth region, and $I^{(g)}=0$ otherwise, and $\{\gamma_1,...,\gamma_G\}$ is the set of additional coefficients associated with each of the G indicator functions. The Tsiatis goodness-of-fit statistic tests the null hypothesis that $\boldsymbol{\gamma}=(\gamma_1,...,\gamma_G)=\mathbf{0}$, and thus the null model,

$$\boldsymbol{\pi}(\mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \tag{3.10}$$

is the best fit to the data out of all of the possible instances of the augmented model. Here, the $\boldsymbol{\gamma}$ are considered the parameters of interest and the $\boldsymbol{\beta}$ are considered nuisance parameters.

The Tsiatis statistic is

$$T = \mathbf{S}'\mathbf{V}^{-}\mathbf{S} \tag{3.11}$$

where $\mathbf{S}$ is a G-dimensional column vector, with general elements

$$S_g = \frac{\partial l}{\partial \gamma_g} \qquad\qquad g=1,...,G \tag{3.12}$$

with l representing the log likelihood, and $\mathbf{V}^{-}$ is any generalized inverse of the $G\times G$ covariance matrix as in (2.48); that is,

$$\mathbf{V} = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}' \tag{3.13}$$

The elements of the matrices on the right-hand side of (3.13), for the gth group and kth covariate, are

$$A_{gg'} = \begin{cases} -\dfrac{\partial^2 l}{\partial\gamma_g\partial\gamma_{g'}} & (g=g';\ g,g'=1,...,G) \\ 0 & (g\neq g') \end{cases} \tag{3.14}$$

$$B_{gk} = \left\{ -\frac{\partial^2 l}{\partial \gamma_g \partial \beta_k} \qquad (g = 1,...,G;\ k = 0,...,K) \right. \tag{3.15}$$

$$C_{kk'} = \left\{ -\frac{\partial^2 l}{\partial \beta_k \partial \beta_{k'}} \qquad (k, k' = 0,...,K) \right. \tag{3.16}$$

Halteman (1980) also derived this statistic in his dissertation, but did not publish his results. His proof, that the rank of $\mathbf{V}$ is $G-1$, is presented in Theorem 5.1.

All of the terms associated with (3.11) are evaluated under the null hypothesis, that $\boldsymbol{\gamma} = \mathbf{0}$ and $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimates of the parameters estimated under the null model, (3.10). Thus, the elements of the vectors and matrices necessary for the calculation of (3.11) are

$$S_g \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=\mathbf{0}} = \sum_{i=1}^{n} (y_i - \hat{\pi}_i) I_i^{(g)} \qquad (g = 1,...,G) \tag{3.17}$$

and

$$A_{gg'} \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=\mathbf{0}} = \begin{cases} \sum_{i=1}^{n} \hat{\pi}_i (1-\hat{\pi}_i) I_i^{(g)} I_i^{(g')} & (g = g';\ g, g' = 1,...,G) \\ 0 & (g \neq g') \end{cases} \tag{3.18}$$

$$B_{gk} \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=\mathbf{0}} = \left\{ \sum_{i=1}^{n} \hat{\pi}_i (1-\hat{\pi}_i) x_{ik} I_i^{(g)} \qquad (g = 1,...,G;\ k = 0,...,K) \right. \tag{3.19}$$

$$C_{kk'} \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=\mathbf{0}} = \left\{ \sum_{i=1}^{n} x_{ik} x_{ik'} \hat{\pi}_i (1-\hat{\pi}_i) \qquad (k, k' = 0,...,K) \right. \tag{3.20}$$

(Tsiatis 1980). Because $\mathbf{V}$ is not full rank (Halteman 1980, Tsiatis 1980), it is necessary to use a generalized inverse of $\mathbf{V}$.

Tsiatis forms groups by partitioning the covariate space. Although partitioning in the covariate space overcomes some of the shortcomings of grouping methods that only look for discrepancies in the direction of the logit, it also suffers from some deficiencies. Tsiatis did not indicate how the partitioning of the covariate space should be accomplished. No methods are specified for determining what number of partitions to use, nor how they should be chosen. Su,

et al. (1991) point out that tests that partition the covariate space, including the Tsiatis test, can draw different conclusions when different partitions are applied. They give several specific examples, one of which shows the Tsiatis statistic, given two choices of partitioning, resulting in p values of 0.04 and 0.38. Lin, Wei and Ying (2002) also point out that "the partition of the covariate space is arbitrary and different partitions may result in conflicting conclusions".

Another potential difficulty presented by partitioning in the covariate space is that the partition can involve as many dimensions as there are covariates. As the number of dimensions increases, the number of groups increases exponentially with the number of covariates, and a compromise has to be reached between creating enough groups for a particular covariate, so that not too much information is lost, and limiting the number of overall groups to a manageable number. For example, if there are three continuous covariates in the logistic model, a coarse partitioning of each into two groups would create eight groups $\left(\text{i.e. } 2^3\right)$, but would potentially result in a loss of a great deal of information about each covariate. If each covariate was partitioned instead into a larger number of levels, more information would be retained, but many more groups would be created. This would mean that a very large number of observations could be required to sufficiently populate the groups.

Halteman showed that the asymptotic distribution of T is not affected by random cells, and thus other methods of forming groups such as the deciles-of-risk method could also be used. Simulation results reported by Halteman (1980) in his dissertation support the conjecture that the distribution of T is approximately $\chi^2(G-1)$ when applied to finite samples, including cases when the deciles-of-risk grouping method was applied. He recommends that when the deciles-of-risk method is applied, T should be compared to $\chi^2(G-1)$ when used to test model fit.

3.2.3 Pigeon-Heyse Goodness-of-Fit Test Statistic

An alternative that combines characteristics of both HL and T is the goodness-of-fit statistic of Pigeon, et al. (1999b), denoted by J^2. To account for the heterogeneity of the predicted

probabilities within partitioned groups (Dreiseitl and Osl 2012), they multiply HL, (3.6), by a "correction term". The form of J^2 is

$$J^2 = \sum_{g=1}^{G} \frac{\left\{ \sum_{i \in \psi_g} \left(y_i - \hat{\pi}_i \right) \right\}^2}{\varphi_g n_g \bar{\pi}_g \left(1 - \bar{\pi}_g \right)} \tag{3.21}$$

where the correction factor is

$$\varphi_g = \frac{\sum_{i \in \psi_g} \hat{\pi}_i \left(1 - \hat{\pi}_i \right)}{n_g \bar{\pi}_g \left(1 - \bar{\pi}_g \right)} \tag{3.22}$$

The J^2 statistic simplifies to

$$J^2 = \sum_{g=1}^{G} \frac{\left\{ \sum_{i \in \psi_g} \left(y_i - \hat{\pi}_i \right) \right\}^2}{\sum_{i \in \psi_g} \hat{\pi}_i \left(1 - \hat{\pi}_i \right)} \tag{3.23}$$

Pigeon, et al. (1999b) establish the approximate asymptotic distribution of J^2 through simulations. In the settings they evaluated, n_g subjects were classified into one of G mutually exclusive outcomes, where $\sum_{g=1}^{G} n_g = n$. They define π_{ig} as a true but unknown probability that subject i is in the gth outcome state. Estimated probabilities were designated by $\hat{\pi}_{ig}$, and $\sum_{i=1}^{n_g} \hat{\pi}_{ig} = E_g$. The observed outcomes for subject i is in the gth outcome state were y_{ig}, and $\sum_{i=1}^{n_g} y_{ig} = O_g$. The first statistic they compared to J^2 was

$$X^2 = \sum_{g=1}^{G} \frac{\left(O_g - E_g \right)^2}{E_g} \tag{3.24}$$

which, when all of the $\hat{\pi}_{ig}$ within the gth outcome states are equal , is Pearson's chi-squared statistic. As noted earlier, the asymptotic distribution is known in this case. They considered situations where the $\hat{\pi}_{ig}$ within the gth outcome state were not all equal. The other statistic they compare to J^2 is the quadratic form

$$Q = (\mathbf{y} - \hat{\boldsymbol{\pi}})' \mathbf{V}^{-} (\mathbf{y} - \hat{\boldsymbol{\pi}}) \tag{3.25}$$

where $\mathbf{y}' = \left[O_1, O_2, ..., O_g\right]$, $\hat{\boldsymbol{\pi}}' = \left[E_1, E_2, ..., E_g\right]$, $\mathbf{V} = \sum_{i=1}^{n} \left\{ Diag\left(\hat{\boldsymbol{\pi}}_i\right) - \hat{\boldsymbol{\pi}}_i \hat{\boldsymbol{\pi}}_i' \right\}$ and is singular, and $\mathbf{V}^{-}$ is a generalized inverse of $\mathbf{V}$. This has an asymptotic distribution that is $\chi^2(G-1)$ (Kendall, Stuart and Ord 1994, Pigeon, et al. 1999b). They state that J^2 is an approximation of Q.

To compare the statistical properties of J^2, X^2 and Q, they simulated data sets by studying all combinations of the following settings: 1) the number of observations ($n = 50$, 100 and 200), 2) the number of distinct outcomes ranging from $G = 2$ to 20 by two's, and 3) the number of distinct estimated probability vectors taking values $0.04n$, $0.1n$, $0.2n$, $0.5n$ and n. Under these settings, the cell boundaries are fixed before the data are collected. They found that under these settings, J^2 has an asymptotic distribution that is approximately $\chi^2(G-1)$.

Pigeon and Heyse state that J^2 is not dependent on the grouping strategies and that grouping strategies other than the deciles-of-risk, such as the strategy of partitioning the covariate space before data collection suggested by Tsiatis, can be used to form groups. An example of a grouping method used with J^2 from the literature is partitioning based on clusters within the data space (Dreiseitl, et al. 2012).

Although Pigeon and Heyse state that the distribution of J^2 does not depend on the grouping method used, their simulations do not appear to involve cases where group boundaries are random, so it is unclear from their simulation study how grouping strategies that create random cell boundaries would affect the distribution of J^2. In their study, Pigeon and Heyse do compare J^2 to HL using the Low Birth Weight data set described in *Applied Logistic Regression* (Hosmer, et al. 2000), as well as to a goodness-of-fit statistic by Bull (1994) - an extension of HL to polychotomous regression models, which also uses the deciles-of-risk method to form groups. They note that the values of HL and J^2 are similar. In this case the groups are formed based on predicted probabilities, and the group boundaries are random.

3.3 Other GOF Statistics for Logistic Models with Continuous Covariates

We now review some other statistics used to assess the fit of a logistic regression models with continuous covariates. Several of these are presented as alternatives to HL. In some cases, simulation studies have been conducted that compare the performance of HL to that of these alternative statistics.

3.3.1 Goodness-of-Fit Statistics with Grouping Methods Based on Clustering

To avoid the problems presented by the deciles-of-risk grouping method, Xie (2005) and Xie, Pendergast and Clarke (2008) applied a partitioning strategy based on clustering in the covariate space to both a Pearson chi-square type statistic with the same form as HL, and a score statistic with the form of T. The clustering method identifies regions within the covariate space where observations are close, as defined using a criterion such as Euclidean or Mahalanobis distance. They point out that this method has the advantage that observations within these groups will have similar covariate profiles. They state that both of their statistics should have asymptotic distributions that are between $\chi^2(G-K-1)$ and $\chi^2(G-1)$, where K is the number of covariate values. They use the rubric $G=10$ if $K<5$, and $G=K+5$ if $K\geq 5$, applied to $df=G-(K/2)-1$ for the Pearson chi-square type statistic, and $df=G-1$, which is the rank of the conditional covariance matrix of the scores. They compare the performance of their statistics to that of the original HL, which uses the deciles-of-risk grouping method. Both HL and their Tsiatis-like score statistic maintained the test size more consistently, while their Pearson chi-square type statistic was conservative. Both of their test statistics had more power than HL to detect departures from a true underlying model.

Dreiseitl, et al. (2012) offered another strategy for overcoming the problem of detecting lack of fit in a region in the covariate space. They used the Pigeon-Heyse statistic (Pigeon, et al. 1999b), J^2, which is reported to have an asymptotic distribution that is chi-squared with $G-1$ degrees-of-freedom, and applied a grouping method based on clustering. Three strategies they

tested were based on 1) clustering with self-organizing maps; 2) clustering with a K – means algorithm; and 3) random assignment of data points to groups. In their simulations, they varied the dimensionality of the data, studying 5, 10 and 20 dimensional data. Their simulation study was small, with data limited to 20 data sets, with only 10 data sets to evaluate the type I and type II error rates of HL and J^2 using the three cluster grouping methods. They also calculated these statistics for a real world data set. They found that J^2 offered only slightly better performance than HL, although with such small samples this result is not very strong. They report that their approach does aid in locating regions of poor calibration in the data space.

3.3.2 Smoothing Methods for Testing the Fit of Logistic Regression

Methods based on non-parametric kernel smoothing offer a strategy that avoids the problems encountered by grouping strategies, particularly those that detect deviations from the model in only the direction of the fitted probabilities, such as the deciles-of-risk method. One goodness-of-fit statistic for logistic regression based on this method was the statistic introduced by le Cessie and van Houwelingen (1991). Their method builds on the approach of Copas (1983), who plotted the non-parametric kernel estimation of model probabilities against the linear predictor to examine model fit graphically, as well as on work by Azzalini, Bowman and Härdle (1989) who generalized Copas' method to compare the function of the null model to a kernel estimate. The le Cessie and van Houwelingen test statistic is based on a kernel estimate of the standardized residuals that has an expectation equal to zero. A smoothing function of these standardized residuals uses the kernel estimate of Nadaraya (1964) and Watson (1964). They describe the function as a weighted average of the residuals in the neighbourhood of a covariate, where bandwidth determines the size of the region over which the residuals are averaged, and the kernel function determines the weighting. Their approach avoids the problems of bias that occurs in methods that directly compare the parametric and non-parametric curves. Although their method avoids the problems of the various grouping methods, it does have the disadvantage that results can depend on the choice of the bandwidth. Hosmer, et al. (1997) compared the le Cessie and van Houwelingen statistic to several other goodness-of-fit statistics for logistic regression, including HL. They applied the uniform kernel weight function for the

"*x* space", as described in the original le Cessie and van Houwelingen paper, as well as a cubic weight for the "*y* space" to the le Cessie and van Houwelingen statistic. Its performance was found to be similar to that of HL at detecting departures from the true model. However, unlike HL, in settings where a cubic weight smooth was chosen and the linear predictor of the model contained three covariates, its Type I error rate was higher than expected in some settings.

3.3.3 Goodness-of-Fit Statistics for Logistic Models with Discrete Covariates

As discussed in section 3.1, X^2 and D may be used to assess the fit of a model to observed data when the number of covariate patterns is fixed. In this case the two statistics are asymptotically equivalent. However, under some circumstances the asymptotic distributions of these statistics can differ. For example, if the number of covariate patterns grows at a similar rate to the overall number of observations, but their ratio remains fixed, then the asymptotic means and variances of X^2 and D may differ (Read and Cressie 1988). This condition is referred to as "sparseness". Which of these two statistics is optimal depends on the particular situation.

Cressie and Read (1984) introduced a family of power-divergence statistics that gives a unifying approach to testing the fit of models with discrete multivariate data. These statistics take the form

$$\frac{2}{\lambda(\lambda+1)}\sum_{j=1}^{J}\text{observed}_j\left[\left(\frac{\text{observed}_i}{\text{expected}_j}\right)^{\lambda}-1\right] \tag{3.26}$$

where J is the fixed number of possible outcomes and λ is a real-valued parameter chosen by the user. They define the two cases that can result in division by zero, that is $\lambda=0$ and $\lambda=-1$, as the limits where $\lambda\to 0$ and $\lambda\to -1$, respectively. Which value of λ gives the optimal test statistic depends on the particular circumstances, such as if there is a condition of sparseness or whether the null hypothesis is true. They suggest (Read, et al. 1988) that a reasonable choice for the value of λ is a value that lies in the range $\lambda\in(-1,2]$. This range includes both X^2, for which $\lambda=1$, and D, for which $\lambda=0$. When certain details of the circumstances are unknown, for example the alternative model, then they suggest a test statistic which lies between X^2 and

D, with $\lambda = 2/3$, as a compromise with excellent properties when the sample size is small (see Read and Cressie, 1988, chapter 5). Read, et al. (1988) report the asymptotic distribution of the power-divergence tests are central chi-squared with degrees of freedom equal to $G(J-1)-K$, when all three parameters are fixed.

In a series of papers, McCullagh(1985, 1986) considered the effect of sparse data on both X^2 and D. Specifically he considers the case when the number of cells is increasing to infinity, rather than the mean count within the cells increasing to infinity. McCullagh argues that the conditional distribution of X^2 and D, rather than their marginal distributions, are relevant for assessing goodness-of-fit of GLMs when the parameters have been estimated with reference to the data, rather than fixed in advance. To remove the distributional dependence of the statistics on the parameter estimates, he conditions on the sufficient statistic for the parameter estimates. He gives an approximate analytical solution for the conditional distributions of both X^2 and D for GLM with canonical link functions. He found that for binary data, D was uninformative as a goodness-of-fit test because it is a function of the sufficient statistic, and when every observation has its own covariate pattern, D is completely independent of the observations. He instead recommends the use of X^2 as a goodness-of-fit test for binary data when data are sparse, and presents a standardized Pearson statistic for goodness-of-fit that is conditional on the sufficient statistic of the unknown parameters. He derives the first three unconditional and conditional moments of X^2, which are necessary for the calculation of his generalized statistic. McCullagh shows, using the first order correction term to X^2 , that X^2 and the sufficient statistic are independent (i.e. they are orthogonal), thus accommodating the estimation of parameters referencing the observed data rather than determining them in advance. A second-order correction is applied using Edgeworth expansion to obtain improved approximations for the distributions of X^2. Alternately, Farrington (1996) suggests a comparison of McCullagh's statistic to a $N(0,1)$.

Osius and Rojek (1992) applied the work of Cressie, et al. (1984), Read, et al. (1988), McCullagh (1985), and McCullagh (1986) to models where the number of possible outcomes is

increasing. They derived a statistic similar to McCullagh's statistic, which when applied to binary data is a score test for the fit of the hypothesized model against a particular enlarged model alternative. Their derivation is based on the calculation of the first two moments of the Cressie-Read power-divergence statistic, and which, under certain conditions, has an asymptotic standard normal distribution.

Pulkstenis and Robinson (2002) suggested two alternative goodness-of-fit statistics that are also intended to overcome the problems created when groups contain subjects with a wide range of values of the covariate. Their two statistics are similar to X^2 and D, but can be applied to logistic regression models containing both categorical and continuous covariates. They use the Hosmer-Lemeshow strategy of grouping data, but also cross-classify categorical variables to allow the structure of the individual covariate patterns to remain intact. First observations are sorted by unique covariate patterns based only on the categorical covariates. Next, within each of the first level groups the observations are sorted by fitted probabilities. Finally, the groups are split again into two, with division at on the median categorical response. If the median response is an actual value, then it is placed in the lower group. Under the null hypothesis, the statistic is reported to have an approximate asymptotic chi-squared distribution of two times the number of unique covariate patterns (based only on the categorical covariates), minus the number of categorical variables in the model, minus two. Pulkstenis, et al. (2002) found that the power of their statistic was greater than that of HL, with regard to its ability to detect the omission of an intercept term from the true model. In addition, their grouping method allows for an analysis of observed and expected cell counts based on the covariate classification, which can aid in identifying poor fit within the covariate space. However, Pulkstenis and Robinson point out that their model is applicable only in certain situations - when a model contains both continuous and categorical covariates, and there are not too many cross-classifications among the categorical covariates in the model. They suggest applying their test in conjunction with HL.

3.3.4 Score Tests for Assessing the Fit of Logistic Regression Models

Stukel (1988) proposed a score test to evaluate the goodness-of-fit test of logistic regression models. Her test is designed to detect both symmetric and asymmetric deviations from the

hypothesized model. She introduces a generalized logistic model containing two additional shape parameters. These enable modification of the tails of the model curve. The model is expressed as

$$\mu(\eta) = \frac{\exp\left(h_\alpha(\eta)\right)}{1+\exp\left(h_\alpha(\eta)\right)} \tag{3.27}$$

where

$$h_\alpha(\eta) = \log\left(\frac{\mu}{1-\mu}\right) \tag{3.28}$$

Both are strictly increasing nonlinear functions of η indexed by two shape parameters, α_1 and α_2. Stukel defines the parameters as follows:

for $\eta \geq 0 (\mu \geq 0.5)$,

$$h_\alpha = \begin{cases} \alpha_1^{-1}\left(\exp\left(\alpha_1 |\eta|\right)-1\right) & \alpha_1 > 0 \\ \eta & \alpha_1 = 0 \\ -\alpha_1^{-1}\left(\log\left(1-\alpha_1 |\eta|\right)\right) & \alpha_1 < 0 \end{cases} \tag{3.29}$$

and when $\eta \leq 0 (\mu \leq 0.5)$,

$$h_\alpha = \begin{cases} -\alpha_2^{-1}\left(\exp\left(\alpha_2 |\eta|\right)-1\right) & \alpha_2 > 0 \\ \eta & \alpha_2 = 0 \\ \alpha_2^{-1}\left(\log\left(1-\alpha_2 |\eta|\right)\right) & \alpha_2 < 0 \end{cases} \tag{3.30}$$

The two shape parameters are independent, allowing the tails to be symmetric or asymmetric. The usual regression coefficients are taken as nuisance parameters, and the hypothesis tested is that the shape parameters are equal to zero, and thus the regular logistic model is a good fit.

Liu, Nelson and Yang (2012) proposed another test statistic, this time based on the strengths of the earlier test statistics by Hosmer, et al. (1980), Tsiatis (1980), Stukel (1988), and Pulkstenis, et al. (2002), which seeks to overcome some of their deficiencies. Liu, et al. (2012) point out that the Pulkstenis-Robinson grouping tests, which require a categorical component to the covariate vector, can make the dimensions of the covariate vector very large. This is particularly

the case if interaction terms between the categorical and continuous covariates are included in the model, which can result in low power to detect lack of fit. Their new approach, like the Tsiatis test, uses an augmented model along with a score test to evaluate the fit of an original model. However, they employ a new grouping strategy. They create the augmented model by partitioning the data into "pseudo replicates" based on scenarios in which discrete covariates are either present or not. If the model contains only continuous covariates, then the partition is based on the fitted probabilities grouped into G quartiles. If the model contains both continuous and discrete covariates, a two-level sub-grouping strategy is applied. First, subgroups are formed based on the distinct covariate patterns of the discrete covariates only. Then a second partitioning is performed within the first subgrouping, with the continuous covariate space split into four parts based on the quartiles of the fitted probabilities. The augmented and original models are then compared to determine the fit of the model being evaluated. A benefit of their method is that the augmented model can inform the user about the nature of the lack of fit, and can provide information that may suggest an improvement to the model. This method may have limited power if the number of data points is small.

Barnhart and Williamson (1998) developed a goodness-of-fit score test similar to that of Tsiatis for correlated binary models with repeated outcome measures, where the covariate space is partitioned by cross-classifying the covariates. However, a large number of groups can result when a model contains both continuous and discrete covariates, and, if the number of observation is not very large, this can result in some of the grouping being sparsely populated.

Horton, et al. (1999) extended the work of Hosmer, et al. (1980), Tsiatis (1980), and Barnhart, et al. (1998) by developing a score test for generalized estimating equations (GEE) models of binary outcomes that are repeatedly measured. To avoid the problems of partitioning the covariate space, they apply the deciles-of-risk grouping method of Hosmer and Lemeshow, forming groups based on estimates of risk obtained using GEE estimator methods. They report that their statistic has an asymptotic distribution that is $\chi^2(G-1)$ under the null hypothesis. They note that if the assumption of independence amongst the outcomes were to be assumed, then their statistic is identical to the Tsiatis but with partitioning based on deciles-of-risk.

Archer, Lemeshow and Hosmer (2007) proposed a goodness-of-fit test for logistic regression that uses the Tsiatis model to construct an F-adjusted Wald test, rather than the usual score test, to test the hypothesis that the coefficients associated with the indicator functions are equal to zero. In this case the model fitted includes the indicator functions. Under this test, the groups are formed based on the deciles-of-risk method rather than on a partition of the covariate space. Because there is linear dependence between the indicator functions and the function 1 associated with the usual intercept term, β_0, they encountered difficulty fitting all of the terms of the model. The computer software would drop one of the indicators for the groups, and thus not all of the deciles would be represented, resulting in loss of power. Because of this problem they do not recommended the use of this method.

Building on their earlier work (le Cessie, et al. 1991) on a goodness-of-fit test for logistic regression based on non-parametric kernel smoothing methods, le Cessie and van Houwelingen (1995) showed that their original method can be used as a score test in a random effects model, and can be extended to GLMs with canonical link functions. Hosmer, et al. (1997) note that a special case of the le Cessie and van Houwelingen statistic is one introduced earlier by Copas (1989), that only considers the numerator of X^2, the unweighted residual sum of squares, $\sum_{i=1}^{n}\left(y_i - m_i\hat{\pi}_i\right)^2$. Both the studies by Hosmer, et al. (1997) and Kuss (2002) included this statistic in their analysis comparing several goodness-of-fit tests for logistic regression models. Hosmer, et al. (1997) also gives a method for calculating its asymptotic moments, and shows how to use it to test for goodness-of-fit of a logistic regression model.

3.4 Studies Comparing the Performance of GOF Statistics for Binary Logistic Regression Models

Several studies compare the performance of HL to those of other GOF test statistics for logistic regression models with continuous covariates. In Hosmer, et al. (1997), the performance of HL is compared to that of several other statistics, including a statistic based on smoothed residuals by le Cessie, et al. (1991) and by Royston (1992), the score test by Stukel (1988), the X^2, and

an unweighted residual sum-of-squares proposed by Copas (1989). They found generally that all of the statistics they studied, except for the Royston statistic, maintained the correct test size. In a few of the settings, however, the rejection percentage was higher than expected, particularly for the le Cessie and van Houwelingen statistic with cubic weights in the "y-space". The statistics with the highest power to detect the omission of a quadratic term were the X^2 and the unweighted residual sum-of-squares. All of the statistics had low power to detect the omission of an interaction term. They all had somewhat more power to detect an incorrect link function, with the Stukel statistic having the most power.

Kuss (2002) conducted another large study comparing several other goodness-of-fit statistics to HL. He viewed the study as supplemental to that of the earlier study by Hosmer, et al. (1997). Several different settings were considered. The design included varying the degrees of sparseness. The other statistics studied included both X^2 and D, as well as statistics by Osius, et al. (1992) (X_O^2), McCullagh (1985) (X_{Mc}^2), and Farrington (1996) (X_F^2). Also included were the residual sum of squares test (RSS) (Copas 1989) , and the information matrix test (IM) from the economics literature (White 1982), which is based on comparing two different estimators of the information matrix that are equal if the model is a good fit. As expected, Kuss found that both X^2 and D generally performed poorly when data was sparse. Specifically, they did not maintain test size while the other test statistics did. Generally the other test statistics performed as well as HL, and in some cases better. For example, the Farrington test generally had more power than HL. They noted that in a real world example, when outliers were removed and the statistics recalculated, the updated result of HL inexplicably indicated that the model fit had worsened, while other test statistics indicated improved fit.

Hosmer and Hjort (2002) compared the performance of HL, X^2, and the unweighted sum-of-squares tests to three statistics that have a weighted statistical process applied. Two of the new test statistics were grouped deciles-of-risk tests. These were weighted versions of HL and X^2. The weighting method used was similar to that of Su, et al. (1991). Their goodness-of-fit techniques include graphical and numeric methods, and are based on the cumulative sums of residuals over certain coordinates. Their strategy was to compute partial sums over successively

larger partitions of the covariate space that are fixed. Hosmer and Hjort instead use the estimated logit to form partitions, and weight the individual residuals of the tests using a weight function that is optimal for an alternative model that is specified by the user. This strategy requires the user to make an educated guess about what terms might be missing. This allows for potentially more insight into specific causes of lack of fit, but also presents the risk that the user may not test for a term that has been omitted from the true model. The simulations indicated that all of the statistics had correct size. No single test had more power than any of the others examined to detect lack of fit in the model due to an omitted covariate or due to an incorrectly specified link function. However, one of the new statistics, referred to as test #12 in their paper and which is based on partial sums-of residuals from the fitted model weighted with a model specific omitted covariate, had more power to detect the omission of an interaction term, which is often difficult to detect.

3.5 Goodness-of-Fit Statistics for Non-Canonical GLM

Although the selection of the canonical link function when building a GLM model, such as the logit link, can offer several advantages (Czado and Munk 2000), including a guarantee of maximum information and a simple interpretation of regression parameters, the estimates of the regression parameters can be biased if the canonical link is in fact incorrectly specified. Czado, et al. (2000) give several examples of situations that warrant the use of a non-canonical link as well as situations that do not, and present a generalized p-function that can be used to aid in the decision. If a non-canonical link is selected, then the problem of assessing the non-canonical model's overall fit to the data still remains.

3.5.1 Statistics to Assess the Fit of Non-Canonical GLM with Discrete Covariates

Both the D and X^2 statistics outlined in the previous section can also be used to assess the fit of non-canonical GLM if the covariates in the model are discrete (Windmeijer 1995). Both are examples of statistics that fall under the Read and Cressie power-divergence statistics. These

statistics are appropriate when the number of observations with any of the unique covariate patterns formed under the model is not small.

Citing the widespread use of GLMs with non-canonical links and a need for the development of goodness-of-fit techniques to evaluate them, Farrington(1995, 1996) follows the ideas of McCullagh (1985, 1986) in deriving a goodness-of-fit statistic for non-canonical GLM models based on the conditional moments of X^2. Like McCullagh, he considers the case when there is extensive discrete covariate data, but $n \to \infty$ rather than the mean count within the cells going to infinity. Based on estimating equations developed by McCullagh (1986), his approach is to embed a canonical GLM into a family of GLMs with arbitrary link function and the addition of first order components. He defines an independent random variable, Y_i, $i = 1,2,...,n$, with mean μ_i and variance function V_i. In this case, $\phi = 1$, and the link function, $g(\cdot)$, is such that

$$g^{-1}(\mu) = \eta = \mathbf{x}'\boldsymbol{\beta} \tag{3.31}$$

Here μ will be denoted as $\pi(\boldsymbol{\beta})$. The estimating equations for obtaining the maximum likelihood estimates of the regression parameters are, for $k = 1,...,K$ covariates,

$$\sum_{i=1}^{n}\left[\frac{y_i - \pi_i(\boldsymbol{\beta})}{V_i}\frac{\partial\left[\pi_i(\boldsymbol{\beta})\right]}{\partial_i \beta_k}\right] = 0 \tag{3.32}$$

The wider family of models has variance ϕV, and a supplementary unbiased estimating equation for estimating ϕ is added,

$$\sum_{i=1}^{n}\left[\frac{\left(y_i - \pi_i(\boldsymbol{\beta})\right)^2}{V_i} - \phi\right] + \sum_{i=1}^{n} a_i\left(y_i - \pi_i(\boldsymbol{\beta})\right) = 0 \tag{3.33}$$

where a_i is a function of μ_i. The statistic of the wider family is

$$\sum_{i=1}^{n}\left[\frac{\left(y_i - \hat{\pi}_i\right)^2}{\hat{V}_i}\right] + \sum_{i=1}^{n} a_i\left(y_i - \hat{\pi}_i\right) \tag{3.34}$$

Farrington notes that the choice of $a_i = 0$ gives the original X^2. Farrington showed that this modified statistic is asymptotically independent of the regression parameter, and is thus an

improvement on the McCullagh method. He shows that a statistic, X_F^2, using a particular choice of correction factor, $a_i = -(\partial V_i / \partial \mu) V_i^{-1}$, induces local orthogonality between the modified statistic and the regression parameters, and has minimal variance within the family. In the logistic setting, where $\hat{\pi}_j$ denotes the predicted probability for the jth covariate pattern calculated, n_j the number of observations with that covariate pattern, and y_j the number of outcomes where $y = 1$, then the statistic is

$$X_F^2 = \sum_{j=1}^{J} \frac{\left(y_j - n_j \hat{\pi}_j\right)^2}{n_j \hat{\pi}_j \left(1 - \hat{\pi}_j\right)} + \sum_{j=1}^{J} \frac{-\left(1 - 2\hat{\pi}_j\right)\left(y_j - n_j \hat{\pi}_j\right)}{n_j \hat{\pi}_j \left(1 - \hat{\pi}_j\right)} \tag{3.35}$$

(Kuss 2002). Farrington uses this choice of correction factor in a standardized statistic that can be used to assess the fit of a model, and that is compared to a standard normal distribution. A deficiency of X_F^2 is that if $n_j \equiv 1$, then $X_F^2 = J$ and the test will never conclude that the model does not fit (Kuss 2002).

Deng and Paul (2000, 2001, 2002, 2011) furthered the work of Farrington (1996). They derived approximations of the first three moments of the unconditional and conditional distributions of the deviance (Deng, et al. 2000) and the first four moments of the unconditional and conditional distributions of the modified X^2 for GLM with non-canonical links for discrete but sparse data (Deng 2001, Paul, et al. 2011). They also derived a score test (Paul, et al. 2002) based on the modified Pearson statistic of Farrington to test for over-dispersion in GLM with sparse discrete data.

3.5.2 Goodness-of-Fit Statistics for Assessing the Fit of Probit Models with Continuous Covariates

In the economics literature, Andrews (1988) extended the Pearson chi-squared statistic to parametric models with covariates. This included a statistic that can be used to test the goodness-of-fit of probit models with continuous covariates. He suggests a variety of "non-parametric" partitioning methods to create cells. That is, the cells are formed using methods that do not rely on the specific conditional parametric model. These methods partition based on both

the outcomes and the covariates, and the cells are considered random. After partitioning, the test statistic is calculated as a quadratic form based on the difference between the number of observed outcomes in each cell and the number expected in each cell, conditional on the observed covariates.

3.5.3 Assessing the Fit of Log Binomial Models

Blizzard and Hosmer used the Hosmer-Lemeshow statistic to assess the fit of both log binomial (Blizzard, et al. 2006) and the log multinomial (Blizzard and Hosmer 2007) models. In the log binomial study (2006), simulations were conducted to compare the performance of HL, a normalized X^2, and an unweighted sum of squares, when assessing the fit of binary log binomial models. They found that the empirical Type I error rates of the normalized X^2 (3.6-11.6 per cent), were mostly within the range expected, but could be high, while those of the unweighted sum of squares ranged from very low to high (0.5-12.1 per cent). The Type I error rates of HL, however, were near the range of values expected, or slightly lower than the expected (3.2-6.1). Blizzard and Hosmer recommended that the "groups minus two" rule for determining degrees of freedom still be used with HL until more extensive simulations are performed to determine whether a reduction in the degrees-of-freedom is warranted. They also conducted simulations to test the power of HL to detect an incorrect logistic regression model that is fitted to data generated from a log binomial setting. This is the situation in which the user applies HL in the logistic setting. All of the statistics had low to moderate power. They did not study the case of a log binomial model fit to data generated from another link, nor did they study the case when a log binomial model was fitted with a term omitted from the true underlying model's linear predictor.

Chapter 4 Comparison of HL, J^2, and T when Assessing the Fit of Logistic Models

4.1 Introduction

When a logistic regression model contains continuous covariates, it is not appropriate to test the fit of the model using the deviance or Pearson's chi-squared goodness-of-fit test statistics. Instead, other methods must be used. The statistics developed by Hosmer, et al. (1980), Pigeon, et al. (1999b), and Tsiatis (1980), as discussed in Chapter 3, are all appropriate goodness-of-fit tests in this case. To our knowledge, no studies have been presented in the literature that compare these statistics algebraically or compare their performances under the same grouping method. We study both here. First we investigate algebraic relationships between the three statistics. We then evaluate their performances using Monte Carlo simulations. Other test statistics discussed in Chapter 3 are not included in this study due to time constraints.

In this study, we apply the deciles-of-risk grouping strategy to all three of the statistics. One benefit of using the same grouping method is that any differences observed can be more directly attributed to the algebraic differences, rather than a combination of algebraic and grouping method differences. We chose to use the deciles-of-risk method for all three statistics for several reasons. First, the method has previously been applied to all three statistics (Halteman 1980, Hosmer, et al. 1980, Pigeon, et al. 1999b), and the results from these studies give approximate asymptotic distributions under this method. Secondly, the deciles-of-risk method is intuitively appealing to users. It is easy to calculate, and available in most commercial software packages. This has made it a widely used grouping method when calculating HL and some other test statistics, including some discussed in Chapter 3. Finally, although some deficiencies have been reported for the deciles-of-risk method (see Chapter 3), there are no alternative methods that do not also present difficulties. Typically, T and J^2 are calculated using a method that partitions the covariate space without reference to the data. However, as discussed in Chapter 3, there are potential difficulties with this method. If the model contains multiple covariates, the sample size may need to be large to adequately populate the potentially substantial number of regions

required. Also, the selection of partitions is non-standardized and subjective. Different partitioning choices can give differing results.

4.2 Algebraic Comparison

4.2.1 Hosmer-Lemeshow Goodness-of-fit Statistic

One of the most often used goodness-of-fit tests developed to address the issues presented when continuous covariates are included in a binary logistic regression model is the Hosmer-Lemeshow test statistic, HL (Hosmer, et al. 1980). It is widely reported in the literature, and its performance has been compared to many other goodness-of-fit statistics for logistic regression, see (Lemeshow and Hosmer 1982, le Cessie, et al. 1991, Hosmer, et al. 1997, Pigeon, et al. 1999b, Kuss 2002, Pulkstenis, et al. 2002, Dreiseitl, et al. 2012). Its common usage, in part, springs from the fact that it is straightforward, easily implemented, and is currently available in most major statistical packages (e.g. STATA, SAS, and SPSS). The form of HL is similar to that of the X^2, but with grouping accomplished by ordering the predicted probabilities and placing them into groups, usually using the deciles-of-risk method. The HL statistic is conditional upon $\hat{\boldsymbol{\beta}}$ and can be expressed as (3.6). Alternatively, HL can be written in matrix form as

$$HL = \mathbf{S}'\bar{\mathbf{A}}^{-1}\mathbf{S} \tag{4.1}$$

where $\mathbf{S}' = \left[\left\{\sum_{i=1}^{n}(y_i - \hat{\pi}_i)I_i^{(1)}\right\},...,\left\{\sum_{i=1}^{n}(y_i - \hat{\pi}_i)I_i^{(G)}\right\}\right]$, and $\bar{\mathbf{A}}$ is a $G \times G$ diagonal matrix containing elements

$$\bar{A}_{gg'} = \begin{cases} n_g\bar{\pi}_g(1-\bar{\pi}_g) & (g = g';\ \text{where g}, g' = 1,...,G) \\ 0 & (g \neq g') \end{cases} \tag{4.1}$$

In this case, $\left\{I_i^{(1)},...,I_i^{(G)}\right\}$ are a set of indicator functions for the ith observation that are defined as $I_i^{(g)} = 1$ when the covariates lie in it the gth group and $I_i^{(g)} = 0$ otherwise.

4.2.2 Pigeon-Heyse Goodness-of-fit Statistic

The J^2 statistic is also conditional on $\hat{\boldsymbol{\beta}}$ and can be expressed as (3.21), which simplifies to (3.23). Note that when the deciles-of-risk method of grouping is used, the numerators of *HL* and J^2 are the same, but their denominators differ. Another expression for J^2 is

$$J^2 = \mathbf{S}'\mathbf{A}^{-1}\mathbf{S} \tag{4.2}$$

where $\mathbf{S}$ is given following (4.1), and the $G \times G$ diagonal matrix $\mathbf{A}$ with elements

$$A_{gg'} = \begin{cases} \sum_{i=1}^{n} \hat{\pi}_i \left(1-\hat{\pi}_i\right) I_i^{(g)} I_i^{(g')} & \left(g = g'; g, g' = 1,...,G\right) \\ 0 & \left(g \neq g'\right) \end{cases} \tag{4.3}$$

is the covariance matrix when the nuisance parameter, in this case $\boldsymbol{\beta}$, is known (Smyth 2003). Expression (4.2) has the same form as the score test described in (2.47). If

$\sum_{i=1}^{n} \hat{\pi}_i \left(1-\hat{\pi}_i\right) I_i^{(g)} I_i^{(g')} = n_g \bar{\pi}_g \left(1-\bar{\pi}_g\right)$, for all g, then *HL* and J^2 are equal.

4.2.3 Tsiatis Goodness-of-Fit Statistic

Tsiatis (1980) developed a goodness-of-fit test statistic for regression models based on Rao's efficient score test (Rao 1948, Rao 2002). The basic idea of the test is to partition the covariate space (i.e. $x-$space) into G distinct regions and then to test that the model fits the data well, against the alternative that the augmented model in any region produces a model that fits the data better. These adjustments are functions of the indicator functions only, and are thus constant within each region. The conditional probability of a successful outcome is given as (3.9). Recall that the Tsiatis statistic is $T = \mathbf{S}'\mathbf{V}^{-}\mathbf{S}$, as in (3.11), where $\mathbf{S}$ is a G-dimensional column vector $\left(\partial l/\partial\gamma_1,...,\partial l/\partial\gamma_G\right)'$ with l representing the log likelihood and $\mathbf{V}^{-}$ any generalized inverse of the $G \times G$ singular covariance matrix, $\mathbf{V} = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}'$. The terms of $\mathbf{S}$ and $\mathbf{V}$ are given in equations (3.14) through (3.16). These are evaluated under the null hypothesis that $\boldsymbol{\gamma} = 0$, and under $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimates under

the null hypothesis. This results in elements (3.18) through (3.20). Note that the matrices (3.18) and (4.3) are equivalent.

To establish the relationship between the three statistics, T may be expressed using any one of five identities presented in Henderson and Searle (1981) (section 4) for the generalized inverse of a singular matrix of the form $\mathbf{A}+\mathbf{B}\left(-\mathbf{C}^{-1}\right)\mathbf{B}'$. If the identity labelled G_3 in Henderson, et al. (1981) is chosen, then

$$T = \mathbf{S}'\left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}'\right)^{-}\mathbf{S}$$

$$= \mathbf{S}'\mathbf{A}^{-1}\mathbf{S} + \Delta$$

$$= J^2 + \Delta \tag{4.4}$$

where

$$\Delta = \mathbf{S}'\left\{\mathbf{A}^{-1}\mathbf{B}\mathbf{C}^{-1}\left(\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{B}'\mathbf{A}^{-1}\mathbf{B}\mathbf{C}^{-1}\right)^{-}\mathbf{C}^{-1}\mathbf{B}'\mathbf{A}^{-1}\right\}\mathbf{S} \tag{4.5}$$

where $\left(\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{B}'\mathbf{A}^{-1}\mathbf{B}\mathbf{C}^{-1}\right)^{-}$ is a generalized inverse of $\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{B}'\mathbf{A}^{-1}\mathbf{B}\mathbf{C}^{-1}$. If $\Delta = 0$, then $J^2 = T$. It is evident that $\Delta = 0$ if $\mathbf{B}'\mathbf{A}^{-1}\mathbf{B} - \mathbf{C} = \mathbf{0}$. The solution to this equation requires that

$$\sum_{g=1}^{G} \frac{\sum_{i=1}^{n} \hat{\pi}_i\left(1-\hat{\pi}_i\right) x_{ik} I_i^{(g)} \left\{\sum_{i=1}^{n} \hat{\pi}_i\left(1-\hat{\pi}_i\right) x_{ik} I_i^{(g)}\right\}}{\sum_{i=1}^{n} \hat{\pi}_i\left(1-\hat{\pi}_i\right) I_i^{(g)} I_i^{(g)}} = \sum_{i=1}^{n} \hat{\pi}_i\left(1-\hat{\pi}_i\right) x_{ik} x_{ik'} \tag{4.6}$$

for each of the combinations of covariates, k, k'. In the general case the problem appears to be intractable. However, equation (4.6) is satisfied when within any group the predicted probabilities are equal, which was noted previously by Pigeon, et al. (1999b). Then $\mathbf{B} = \mathbf{0}$, and the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are orthogonal (Smyth 2003). In this case, HL, J^2 and T all reduce to X^2 and have an asymptotic distribution that is $\chi^2\left(G-K-1\right)$. The asymptotic distribution of HL was shown by Hosmer, et al. (1980) to follow $\chi^2\left(G-K-1\right)\right) + \sum_{k=1}^{K+1} \lambda_k \chi_k^2\left(1\right)$, where λ_k represents the k-th eigenvalue of the matrix they refer to as $\Sigma\left(\hat{H}^*\right)$, and which is described in

their paper. Through simulations, Hosmer, et al. (1980) found that $\sum_{k=1}^{K+1}\lambda_k\chi_k^2(1) \approx \chi_k^2(K-1)$, and hence $HL \sim \chi^2(G-2)$. However when the predicted probabilities within each of the groups are the same, then $\lambda_k = 0$ for all k, resulting in the changed distribution (Hosmer, et al. 1980, Kendall, et al. 1999). The J^2 statistic is an approximation of the statistic Q, (3.25), described by Pigeon, et al. (1999b). A special case of Q occurs when all of the elements within each group are equal; this is X^2, which has a known asymptotic distribution of $\chi^2(G-K-1))$ (Kendall, et al. 1994, Pigeon, et al. 1999b).

4.3 $HL \leq J^2$

The expressions for HL and J^2 are given in (3.6) and (3.23) respectively. Consider these expressions for the gth group containing n_g ordered predicted probabilities, $\hat{\pi}_1,...,\hat{\pi}_{n_g}$. The contribution of the gth group to each statistic can be expressed as

$$HL_g = \frac{\left\{\sum_{i\in\psi_g}(y_i - \hat{\pi}_i)\right\}^2}{n_g\bar{\pi}_g(1-\bar{\pi}_g)} \tag{4.7}$$

and

$$J^2{}_g = \frac{\left\{\sum_{i\in\psi_g}(y_i - \hat{\pi}_i)\right\}^2}{\sum_{i\in\psi_g}\hat{\pi}_i(1-\hat{\pi}_i)} \tag{4.8}$$

The numerators of (4.7) and (4.8) are the same but the denominators differ. Let D_{HLg} denote the denominator of HL, and D_{J2g} the denominator of J^2, for the gth group. That is,

$$D_{HLg} = n_g\left(\frac{\hat{\pi}_1 + ... + \hat{\pi}_{n_g}}{n_g}\right)\left(1 - \frac{\hat{\pi}_1 + ... + \hat{\pi}_{n_g}}{n_g}\right)$$

$$= \left(\hat{\pi}_1 + ... + \hat{\pi}_{n_g}\right) - \left(\frac{\left(\hat{\pi}_1^{\,2} + ... + \hat{\pi}_{n_g}^{\,2}\right) + \sum_{j=i+1}^{n_g} \sum_{i=1}^{n_g-1} \left(\hat{\pi}_i + \hat{\pi}_j\right)^2}{n_g} \right) \tag{4.9}$$

and

$$D_{J2g} = \hat{\pi}_1 - \hat{\pi}_1^{\,2} + ... + \hat{\pi}_{n_g} - \hat{\pi}_{n_g}^{\,2}. \tag{4.10}$$

Theorem 4.1 *HL*=J^2 if and only if all of the predicted probabilities within each group are equal.

Proof

1. The "if" case.

Assume that all of the predicted probabilities within each group are equal. Show that $HL = J^2$.

Let all of the predicted probabilities within the gth group be equal. Then

$$D_{HLg} = n_g\left(\hat{\pi}_1\right) - \frac{n_g\left(\hat{\pi}_1^{\,2}\right) + \frac{n_g\left(n_g - 1\right)}{2}\left(2\hat{\pi}_1^{\,2}\right)}{n_g}$$

$$= n_g\hat{\pi}_1\left(1 - \hat{\pi}_1\right) \tag{4.11}$$

and

$$D_{J2g} = n_g\hat{\pi}_1\left(1 - \hat{\pi}_1\right) \tag{4.12}$$

and thus $D_{HLg} = D_{J2g}$. It follows that, if for every group the elements within that group are equal, then $HL = J^2$.

2. The "only if" case.

Assume that $HL = J^2$. Show that all of the predicted probabilities within each group are equal.

Given that $HL = J^2$, then $J^2 - HL = 0$, and thus

$$\sum_{g=1}^{G} \frac{\left\{\sum_{i\in\psi_g}\left(y_i - \hat{\pi}_i\right)\right\}^2 \left\{n_g\overline{\pi}_g\left(1 - \overline{\pi}_g\right) - \sum_{i\in\psi_g}\hat{\pi}_i\left(1 - \hat{\pi}_i\right)\right\}}{n_g\overline{\pi}_g\left(1 - \overline{\pi}_g\right)\left\{\sum_{i\in\psi_g}\hat{\pi}_i\left(1 - \hat{\pi}_i\right)\right\}} = 0. \tag{4.13}$$

A trivial case when (4.13) is true, is when all $y_i = \hat{\pi}_i$. The other case is when

$$\sum_{i \in \psi_g} \hat{\pi}_i \left(1 - \hat{\pi}_i\right) - n_g \bar{\pi}_g \left(1 - \bar{\pi}_g\right) = 0. \tag{4.14}$$

The left hand side of (4.14) reduces to

$$\sum_{j=i+j}^{n_g} \sum_{i=1}^{n_g - 1} \frac{\left(\hat{\pi}_i - \hat{\pi}_j\right)^2}{n_g}. \tag{4.15}$$

Since the summands are all non-negative, the only case where (4.15) is equal to 0 is when $\hat{\pi}_i = \hat{\pi}_j$, for all i and j. Thus if $HL = J^2$, then all of the predicted probabilities within each group must be equal.

Theorem 4.2 If any of the predicted probabilities within a group differ, then $HL < J^2$.
Proof

If any two predicted probabilities are unequal, then

$$\sum_{j=i+j}^{n_g} \sum_{i=1}^{n_g - 1} \frac{\left(\hat{\pi}_i - \hat{\pi}_j\right)^2}{n_g} > 0 \tag{4.16}$$

and thus the left-hand side of (4.14) is greater than 0. It follows then that, except in the trivial case when all $y_i = \hat{\pi}_i$, (4.13) is positive and thus $HL < J^2$.

4.4 J^2 Can Be Much Larger Than *HL*

From

Theorem 4.1 and Theorem 4.2, it is known that $J^2 - HL \geq 0$. It can be shown that when the difference between two consecutive predicted probabilities within a group approaches 1, then $J^2 - HL$ will approach infinity. Using the notation of section 4.3, for the gth group, write the difference as

$$J^2{}_g - HL_g = \frac{\left\{\sum_{i\in\psi_g}\left(y_i-\hat{\pi}_i\right)\right\}^2}{\sum_{i=1}^{n_g}\hat{\pi}_i\left(1-\hat{\pi}_i\right)} - \frac{\left\{\sum_{i\in\psi_g}\left(y_i-\hat{\pi}_i\right)\right\}^2}{n_g\dfrac{\sum_{i=1}^{n_g}\hat{\pi}_i}{n_g}\left(1-\dfrac{\sum_{i=1}^{n_g}\hat{\pi}_i}{n_g}\right)} \tag{4.17}$$

Let $\hat{\pi}_j$ and $\hat{\pi}_{j+1}$ represent consecutive elements in the gth group. Then as $\hat{\pi}_j \to 0$, all $\hat{\pi}_k \to 0$ when $k < j$. Likewise, as $\hat{\pi}_{j+1} \to 1$, all $\hat{\pi}_k \to 0$ when $k > j+1$. Therefore, in this case, for all summands of the denominator of $J^2{}_g$, $\hat{\pi}_i\left(1-\hat{\pi}_i\right) \to 0$, and thus $J^2 \to \infty$.

However, as $\hat{\pi}_j \to 0$ and $\hat{\pi}_{j+1} \to 1$

$$HL_g \to \frac{\left[\left(y_1-0\right)+\ldots+\left(y_j-0\right)+\left\{y_{j+1}-1\right\}+\ldots+\left\{y_{n_g}-1\right\}\right]^2}{\left(n_g-j\right)\left(1-\dfrac{\left(n_g-j\right)}{n_g}\right)} \tag{4.18}$$

which is finite as long as $j \neq n_g$ and $j \neq 0$. These are both trivial cases when either all $\hat{\pi}_i = 0$ or all $\hat{\pi}_i = 1$. Thus it follows that as $\hat{\pi}_j \to 0$ and $\hat{\pi}_{j+1} \to 1$, $J^2 - HL \to \infty$.

In practice, a situation where the difference between two consecutive predicted probabilities, $\hat{\pi}_{j+1} - \hat{\pi}_j$, might approach 1 is when the covariate values have a bimodal distribution. The difference is maximized when there are equal numbers of covariate values in the subgroups.

4.5 Simulation Study Comparing *HL*, J^2 and *T*

Simulation studies were conducted to verify the reported asymptotic distributions of HL, J^2 and T when the deciles-of-risk grouping method is used, as well as to compare the performances of the three statistics. The studies examined how well each statistic controlled the Type I error rate when a correctly specified (null) model was fitted to generated data, and compared the power of each statistic to detect a departure from a true underlying model. An analysis was also made of how the decision to reject or to not reject the null hypothesis differed

among the test statistics. We refer to this as "decision agreement". That is, even if the power of the test statistics were similar, did their decision to reject the null hypothesis differ among individual samples. Several settings were considered, with the following allowed to vary: 1) sample size $(n = 100, n = 500)$; 2) number and characteristics of the covariates in the model; and 3) the way in which the fitted model departed from the true underlying model in the power settings (the omission of a quadratic term, the omission of a dichotomous term and an interaction term, and the incorrect specification of the model link function). In order to verify the null distribution of the three statistics, a large number of simulations were conducted in two settings to produce highly accurate results. When comparing the null rejection percentages of the statistics, a smaller number of simulations were conducted under a broader variety of settings. The specific settings follow methods described in Hosmer, et al. (1997), as well as methods used by Xie, et al. (2008). This allows for the comparison of our results to those previously reported for HL and several other goodness-of-fit test statistics. The settings chosen are representative of those encountered in practice, and produce predicted probabilities with a variety of ranges and distributional characteristics. First, a general description is given of how the simulations were conducted, followed with specific details for the settings studied.

4.5.1 Simulation Methods

4.5.1.1 General Simulation Methods

In each simulation, a linear predictor, link function, coefficients, and joint distribution of covariates were chosen in accordance with a true underlying model. A random sample of n covariate vectors were generated for each of r replications of the simulation using the specified distributions and ranges. Probabilities were then generated using the appropriate link function. Binary outcomes, y, were generated by comparing the probabilities to a value u where $u \sim U(0,1)$ according to the rule $y = I(u < \pi)$, where I is an indicator function such that $I = 1$ when the argument is true, and $I = 0$ otherwise. Finally, a specified model was fit to the generated $(\mathbf{x}_i, y_i)$ data, $\hat{\pi}$ was estimated, and values for the HL, J^2 and T statistics were

calculated using the deciles-of-risk grouping method $(G = 10)$. All computer simulations described in this chapter were performed using Stata 10 (StataCorp 2007).

4.5.1.2 Methods to Investigate the Null Distributions of *HL*, J^2, and *T*

Null simulations were conducted, using settings 1 and 5 of Table 4.1, to investigate the reported asymptotic distributions of HL, J^2 and T, when the deciles-of-risk grouping method is used. The distributional characteristics of the model probabilities are also given in Table 4.1. A high replication rate $(r = 100,000)$ was selected to give estimates of the distribution of each statistic. We expected the difference between HL and J^2 to be smallest when the denominators of the statistics were most similar. Under the setting 1, where $x_1 \sim U(-1,1)$, $\beta_0 = 0$, and $\beta_1 = 0.8$, the probabilities are clustered near 0.5, and any differences among the $\hat{\pi}$ within groups would be small. We expected this setting to produce relatively small differences between HL and J^2. Under setting 5, where $x_1 \sim \chi^2(4)$, $\beta_0 = -4.9$, and $\beta_1 = 0.65$, the predicted probabilities are distributed across the $(0,1)$ range and are right-skewed; and thus there is a greater potential for the difference between two consecutive probabilities within a group to be nearer to 1. In this case, based on the results discussed in section 4.4, we expected that the differences between HL and J^2 would be larger. Based on the reported asymptotic distributions of the three statistics (Halteman 1980, Hosmer, et al. 1980, Tsiatis 1980, Pigeon, et al. 1999b), and on preliminary results, the value of HL was compared to $\chi^2(8)$, T was compared to $\chi^2(9)$, and J^2 was compared to both $\chi^2(8)$ and $\chi^2(9)$ in these simulations.

Table 4.1 Settings used to examine the null distributions and the power of *HL*, J^2, and *T*.

		Distribution of covariate			Regression coefficients				Distribution Characteristics†				
Setting§	Linear predictor	x_1	x_2	x_3	β_0	β_1	β_2	β_3	$\pi(1)$	*Q1*	*Q2*	*Q3*	$\pi(n)$
1	$\beta_0+\beta_1x_1$	U(-1,1)	.	.	0	0.8	.	.	0.31	0.40	0.50	0.61	0.69
2	$\beta_0+\beta_1x_1$	U(-3,3)	.	.	0	0.8	.	.	0.08	0.23	0.50	0.77	0.92
3	$\beta_0+\beta_1x_1$	U(-4.5,4.5)	.	.	0	0.8	.	.	0.03	0.14	0.50	0.86	0.97
4	$\beta_0+\beta_1x_1$	U(-6,6)	.	.	0	0.8	.	.	0.01	0.08	0.50	0.92	0.99
5	$\beta_0+\beta_1x_1$	$\chi^2(4)$	.	.	-4.9	0.65	.	.	0.01	0.03	0.06	0.20	0.98
6	$\beta_0+\beta_1x_1$	N(0,1.5)	.	.	0	0.8	.	.	0.06	0.31	0.50	0.69	0.94
7	$\beta_0+\beta_1x_1+\beta_2x_2$	Ber(0.5)	U($2x_1$-6,$2x_1$+2)	.	0	0.69	0.1	.	0.40	0.47	0.56	0.63	0.70
8	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$	U(-6,6)	U(-6,6)	U(-6,6)	0	0.27	0.27	0.27	0.03	0.22	0.47	0.76	0.97
9	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$	N(0,1.5)	N(0,1.5)	N(0,1.5)	0	0.27	0.27	0.27	0.16	0.4	0.5	0.61	0.84
10	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$	U(-6,6)	N(0,1.5)	$\chi^2(4)$	-1.3	0.27	0.27	0.22	0.05	0.22	0.42	0.62	0.93
11	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2	.	-1.1	1.3	0.0	.	0.01	0.06	0.26	0.76	0.96
12	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2	.	-2.0	1.0	0.2	.	0.04	0.05	0.14	0.58	0.94
13	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2	.	-2.3	0.9	0.3	.	0.06	0.07	0.11	0.49	0.94
14	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2	.	-2.7	0.7	0.4	.	0.06	0.07	0.12	0.40	0.93
15	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2	.	-3.2	0.6	0.5	.	0.04	0.06	0.13	0.36	0.93

§ All of the settings were used to examine the adequacy of the reported null distributions of *HL*, J^2 and *T* when the deciles-of-risk grouping method was applied. Only settings 11-24 were used to compare the power of *HL*, J^2 and *T* to detect incorrectly specified models.

† Expected values of the smallest, largest, and three quartiles of the resulting distribution of the logistic probabilities for a sample size of 500.

Table 4.1 Settings used to examine the null distributions and the power of *HL*, J^2, and *T*. (cont.)

		Distribution of covariate			Regression coefficients				Distribution Characteristics				
Setting	Linear predictor	x_1	x_2	x_3	β_0	β_1	β_2	β_3	$\pi(1)$	$Q1$	$Q2$	$Q3$	$\pi(n)$
16	$\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3$	U(-3,3)	x_1^2	U(-3,3)	-1.1	1.3	0.0	1.0	0.00	0.05	0.29	0.70	0.99
17	$\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3$	U(-3,3)	x_1^2	U(-3,3)	-2.0	1.0	0.2	1.0	0.00	0.04	0.23	0.57	0.99
18	$\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3$	U(-3,3)	x_1^2	U(-3,3)	-2.3	0.9	0.3	1.0	0.01	0.04	0.21	0.52	0.99
19	$\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3$	U(-3,3)	x_1^2	U(-3,3)	-2.7	0.7	0.4	1.0	0.00	0.04	0.18	0.49	0.99
20	$\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3$	U(-3,3)	x_1^2	U(-3,3)	-3.2	0.6	0.5	1.0	0.00	0.03	0.14	0.42	0.99
21	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	U(-3,3)	Ber(0.5)	.	-1.8	0.1	0.3	0.1	0.08	0.14	0.16	0.19	0.31
22	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	U(-3,3)	Ber(0.5)	.	-1.8	0.1	0.7	0.2	0.07	0.15	0.17	0.23	0.53
23	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	U(-3,3)	Ber(0.5)	.	-1.8	0.1	1.1	0.4	0.08	0.15	0.17	0.29	0.64
24	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	U(-3,3)	Ber(0.5)	.	-1.8	0.1	1.8	0.6	0.07	0.15	0.18	0.42	0.87

Table 4.2 The linear predictors and distributional characteristics of the Stukel models used in the power simulations.

		Linear						Distribution Characteristics				
Setting	Model	Predictor	x_1	α_1	α_2	β_0	β_1	$\pi(1)$	$Q1$	$Q2$	$Q3$	$\pi(n)$
25	Probit	$\beta_0+\beta_1x_1$	U(-3,3)	0.165	0.165	0	0.8	0.06	0.20	0.51	0.82	0.94
26	Long tail	$\beta_0+\beta_1x_1$	U(-3,3)	-1	-1	0	0.8	0.20	0.34	0.52	0.70	0.82
27	Short tail	$\beta_0+\beta_1x_1$	U(-3,3)	1	1	0	0.8	0.01	0.08	0.49	0.93	0.99
28	Complementary log-log	$\beta_0+\beta_1x_1$	U(-3,3)	0.62	0.37	0	0.8	0.04	0.17	0.54	0.88	0.97
29	Asymmetric long-short tails	$\beta_0+\beta_1x_1$	U(-3,3)	-1	1	0	0.8	0.04	0.13	0.38	0.73	0.90

4.5.1.3 Methods for Comparing of the Null Empirical Rejection Percentages of *HL*, J^2, and *T*

Fewer replications $(r = 10{,}000)$ were performed to examine how well each of the statistics controlled a Type I error rate at the $\alpha = 0.05$ level. Data simulated under the null hypothesis were conducted in a variety of settings following the methods of Hosmer, et al. (1997) for comparability. The settings used for the linear predictors and their distributional characteristics are given in Table 4.1. The covariate and outcome values were generated, a correctly specified logistic model was fit to the data, probabilities were estimated, and HL, J^2 and T were calculated. Settings were chosen to magnify the potential difference between HL and J^2. It was thought that the magnitude of these differences may affect rejection percentages. As noted above, we expected the differences between HL and J^2 to be relatively small under setting 1. Under setting 4, when $x_1 \sim U(-6,6)$, $\beta_0 = 0$, and $\beta_1 = 0.8$, the probabilities are distributed across the $(0,1)$ range, and concentrated at the extremes. Like setting 5, the differences between HL and J^2 in this case could be larger. Setting 7 was not used in the original Hosmer, et al. (1997) study . It is similar to a setting used to evaluate HL in the log binomial setting in Blizzard, et al. (2006). This setting contains a dichotomous covariate and a uniformly distributed continuous covariate that is associated with the dichotomous covariate. Based on the results presented in Blizzard, et al. (2006), we investigated whether the inclusion of a dichotomous term in the linear predictor of a model may affect the distribution of HL. Settings 8 - 10 produced predicted probabilities that are fairly evenly distributed with moderate to large ranges, which would be likely to produce moderate differences between HL and J^2. The other settings in Table 4.1 were used both for assessing the adequacy of the reported null distributions, and for investigating the power of each statistic to determine whether the Type I error rates were controlled in these settings. These settings are described in section 4.5.1.4.

4.5.1.4 Methods for Comparing the Power of *HL*, J^2, and *T*

A comparison was made of the power of HL, J^2 and T to detect departures from a true underlying model when terms in the true model were omitted or the link function was incorrectly specified. All power simulations were replicated $r = 10,000$ times.

First, in two series of settings (11-15, 16-20), a quadratic term was omitted from the fitted model. The linear predictor and distributional characteristics of the settings for the true underlying model are given in Table 4.1. The logit function (2.26) was used in both the true model used to generate the data and the deficient model fitted to the data. Covariate values were independently generated such that $x_1 \sim U(-3,3)$ and $x_2 \sim U(-3,3)$. The first series uses settings 11-15. Outcome values were generated with the linear predictor, $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$. Regression coefficients were chosen such that the model curve passed through the points $\pi(-1.5) = 0.05$, $\pi(3) = 0.95$, and $\pi(-3) = W$, where $W = 0.01,\ 0.05,\ 0.1,\ 0.2,$ and 0.4. As the value of W increases, the departure from linearity increases. The second series of models, settings 16-20, are similar, but the true model included the additional term, x_2, so that the linear predictor for the series of models is $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + x_2$. A logistic model with linear predictor $\eta = \beta_0 + \beta_1 x_1$ was incorrectly fit to data from the first series, while the linear predictor $\eta = \beta_0 + \beta_1 x_1 + \beta_3 x_2$ was fitted to the data of the second series.

Next, four settings, (21-24), were used to investigate the power of each statistic to detect the omission of both a dichotomous covariate and an interaction term. Again, (2.26) was used to produce both the true and fitted models. Values of a continuous covariate x_1 were generated such that $x_1 \sim U(-3,3)$. A random value $u \sim U(0,1)$ was generated, and the expression $x_2 = I(u > 0.5)$ evaluated. Here I is an indicator function. Additionally an interaction term, $x_3 = x_1 x_2$, was created. Outcome values were generated using the linear predictor $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and regression coefficients chosen such that the model curve passed through the points $\pi(-3,0) = 0.1$, $\pi(-3,1) = 0.1$, $\pi(3,0) = 0.2$, and $\pi(3,1) = 0.2 + V$, where $V = 0.1,\ 0.3,\ 0.5,$ and 0.7. The influence of the interaction term increases as V increases. An

incorrectly specified logistic model with linear predictor $\eta = \beta_0 + \beta_1 x_1$ was incorrectly fitted to the data.

Finally, five settings (25-29), were used to investigate the power of each statistic to detect an incorrectly specified link function. The settings and distributional characteristic of the model probabilities are given in Table 4.2. Data were generated using a link function based on the Stukel generalized logistic model (Stukel 1988) specified in (3.27), rather than the usual logistic model. The π were calculated using a linear predictor with shaping terms, α_1 and α_2, defined in (3.29) and (3.30) respectively. Values of the covariate were generated such that $x_1 \sim U(-3,3)$. The values of α_1 and α_2 used to investigate the power of the three statistics are given in Table 4.2. Two settings (25 and 28), produce models with shapes similar to the probit $(\alpha_1 = \alpha_2 = 0.165)$ and the complementary log-log ($\alpha_1 = 0.62$, $\alpha_2 = 0.37$) models respectively. The other settings produced models with varying tail shapes. Setting 26 produces long tails $(\alpha_1 = \alpha_2 = -1)$, setting 27 produces short tails $(\alpha_1 = \alpha_2 = 1)$, and setting 29 produces asymmetric tails that are long on one side and short on the other ($\alpha_1 = -1$, $\alpha_2 = 1$). A logistic model with the linear predictor $\eta = \beta_0 + \beta_1 x_1$ was incorrectly fitted to the data produced using these Stukel models.

4.5.2 Simulation Results

4.5.2.1 Distribution of J^2

Histograms are presented in Figure 4.1 and Figure 4.2, and summary statistics in Table 4.3. In setting 1, when $x_1 \sim U(-1,1)$, the histograms of the HL (blue) and J^2 (black dashed) values have distributions that follow the probability density function of $\chi^2(8)$ (red) closely. Similarly, the histogram of the T (grey) values follows the probability density function of $\chi^2(9)$ (dash and dot) closely. The Kolmogorov-Smirnov tests evaluate the null hypothesis that the probability distributions of the three statistics are the same as their theoretical and/or reported distributions, against the alternative hypothesis that the probability distributions differ in some way. Our results indicate that the probability distributions of both HL and J^2 were

significantly different from $\chi^2(8)$, and that the distributions of J^2 and T were significantly different from $\chi^2(9)$. Note however, that with the very large samples sizes, the test statistics may be overpowered, and the significant differences may be considered unimportant. In the case of J^2, the largest differences observed, in parentheses, were much higher when $\chi^2(9)$ was assumed to be the theoretical distribution instead of $\chi^2(8)$. The means of all three statistics were slightly higher than the mean of these distributions while the variances were slightly lower. However, the mean and variance of J^2 was much lower than that of $\chi^2(9)$. If it was assumed that $J^2 \sim \chi^2(8)$, all three statistics had the expected percentage of observations above the 90th, 95th and 99th percentiles. On the other hand, J^2 had a much lower proportion of observations than expected if $J^2 \sim \chi^2(9)$.

In setting 5, when $x_1 \sim \chi^2(4)$, the histograms still followed these same chi square density functions, but not quite as closely. All had higher peaks and narrower spreads. The Kolmogorov-Smirnov tests again indicated that the probability distributions of HL and J^2 were significantly different than $\chi^2(8)$, and that the distributions of J^2 and T were significantly different than $\chi^2(9)$. The largest differences observed were again when J^2 was assumed to follow a $\chi^2(9)$ distribution rather than $\chi^2(8)$. Again, the means and variances of HL and T were within the 95% confidence intervals for those of $\chi^2(8)$ and $\chi^2(9)$ respectively. However, this time the mean and variance of J^2 were slightly higher than the values expected if $J^2 \sim \chi^2(8)$. In contrast to the first simulations, all three statistics had less than the expected percentage of observations above the 90th percentile and more than expected above the 99th percentiles for these distributions, again falling outside of 95% confidence intervals for the respective distributions. The percentage of observations above the 95th percentile fell within 95% confidence intervals for all three statistics. If it was assumed that $J^2 \sim \chi^2(9)$, the percentage of observations above all three percentiles was much lower than expected.

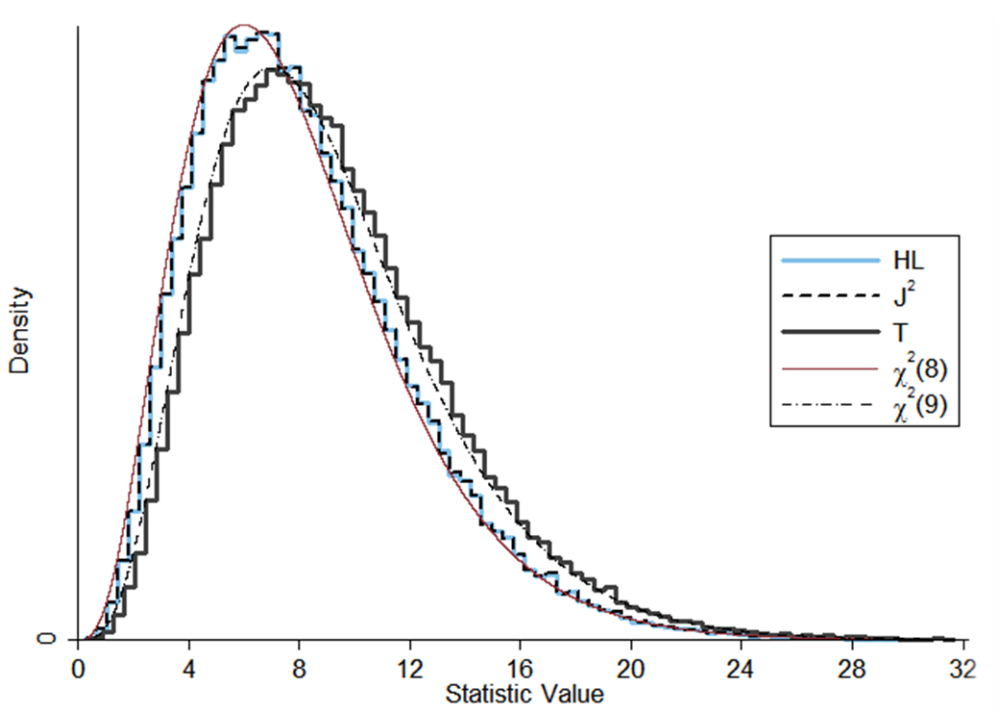

Figure 4.1 Histogram of 100,000 replications of setting 1. (η=0.8x_1, x_1~U(-1,1), n=500) The probability density function curves for $\chi 2(8)$ and $\chi 2(9)$ are included for comparison.

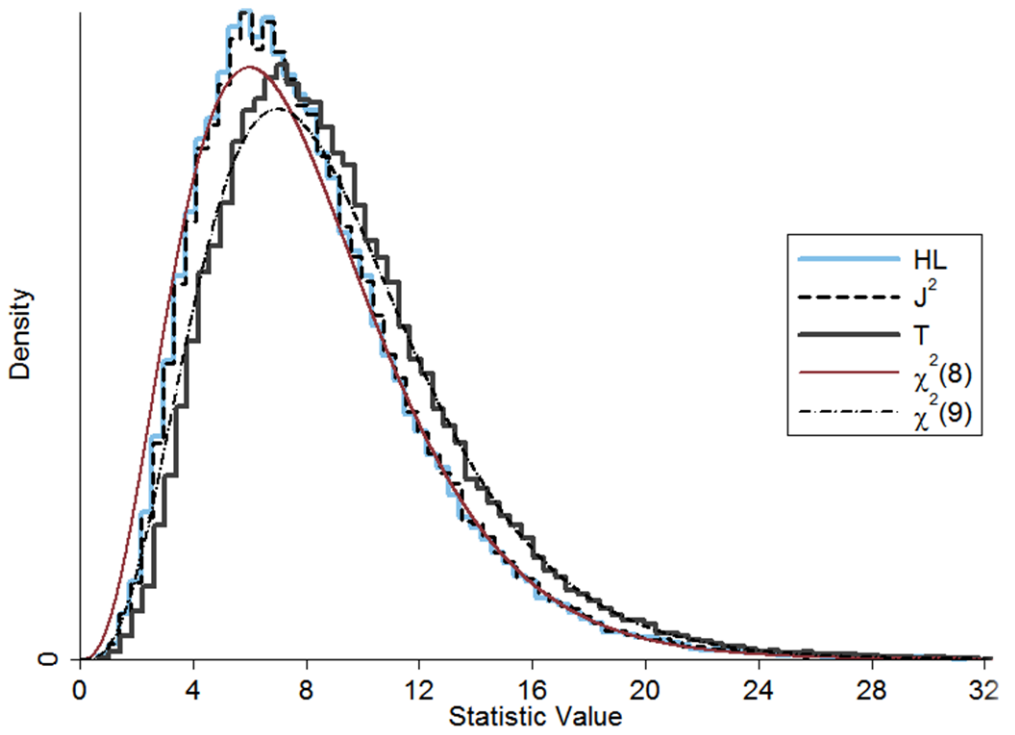

Figure 4.2 Histogram of 100,000 replications of setting 5. (η=-4.9+0.65x_1, x_1~$\chi^2(4)$, n=500) The probability density function curves for $\chi 2(8)$ and $\chi 2(9)$ are included for comparison.

Table 4.3 Summary statistics, rejection percentages, and Kolmogorov-Smirnov test results for *HL*, J^2 and *T* Simulations are based on settings in Table 4.1, using the deciles-of-risk grouping method (G =10) and r =100,000. *HL* and J^2 were compared to critical values for $\chi^2(8)$, and *T* and J^2 to $\chi^2(9)$, when α=0.01, α=0.05, and α=0.1.

				Percent above Percentile**			Kolmogorov-Smirnov†
Setting	Statistic	Mean*	Var.	90th	95th	99th	p-value (D)
1	*HL*	8.03	15.73	9.99	4.86	0.94	0.000 (0.0074)
	J^2_{G-2}	8.04	15.75	10.02	4.87	0.94	0.000 (0.0077)
	J^2_{G-1}	8.04	15.75	6.50	3.01	0.54	0.000 (0.0951)
	T	9.03	17.74	9.99	4.90	0.96	0.000 (0.0065)
5	*HL*	7.98	15.92	9.41	4.88	1.22	0.000(0.0158)
	J^2_{G-2}	8.06	16.22	9.77	5.12	1.29	0.000 (0.0184)
	J^2_{G-1}	8.06	16.22	6.61	3.33	0.79	0.000 (0.1050)
	T	8.99	18.00	9.69	4.94	1.16	0.000 (0.0131)

* Significantly different from χ2(8) at the 5 per cent level if outside interval (7.975, 8.025), and from χ2(9) if it fell outside interval (8.974, 9.026).

**Significantly different from the percentages expected if outside of the 95% confidence intervals (9.814, 10.186), (4.865, 5.135), and (0.938, 1.062).

† H0: No difference between the distribution of the statistic and χ2(df)

HA: The distribution of the statistic and χ2(df) are different.

Our results indicate that both HL and T have distributions near to those reported by others (i.e. $HL \sim \chi^2(G-2)$, $T \sim \chi^2(G-1)$). In the case of J^2 however, our results strongly indicated that the distribution was closer to $\chi^2(G-2)$ rather than $\chi^2(G-1)$ when the deciles-of-risk grouping method is used. Based on these findings, we compared J^2 to the critical value for $\chi^2(8)$ in all of our other analyses.

4.5.2.2 Empirical Rejection Percentage Under the Null Hypothesis

Table 4.4 contains the simulated rejection percentages $(\alpha = 0.05)$ of the settings in Table 4.1 when a correct model is fit to the data generated. Rejection percentages were significantly different from five per cent at the $\alpha = 0.05$ level if they fell outside of the interval (4.6, 5.4). Empirical rejection percentages below, within, and above this criterion were observed for all three statistics, but T was within sampling variation more often (46% for $n = 100$, 88% for

Table 4.4 Simulated null rejection per cent† (*n*=100 and 500, r=10,000, α=0.05) for settings 1-24.

			HL		J^{2}*		*T*	
Setting	Linear predictor	Covariate distributions§	100	500	100	500	100	500
1	$\beta_0+\beta_1 x_1$	x_1~U(-1,1)	5.4	4.8	5.4	4.8	5.2	4.9
2	$\beta_0+\beta_1 x_1$	x_1~U(-3,3)	**4.4**	4.8	**4.5**	4.9	**6.2**	5.0
3	$\beta_0+\beta_1 x_1$	x_1~U(-4.5, 4.5)	4.6	4.4	4.8	**4.5**	**4.4**	4.8
4	$\beta_0+\beta_1 x_1$	x_1~U(-6,6)	**5.9**	4.9	**6.0**	5.0	**5.8**	5.1
5	$\beta_0+\beta_1 x_1$	x_1~$\chi^2(4)$	**6.0**	4.9	**6.2**	5.1	**5.7**	5.0
6	$\beta_0+\beta_1 x_1$	x_1~N(0,1.5)	4.7	4.6	4.9	4.8	4.7	4.6
7	$\beta_0+\beta_1 x_2+\beta_2 x_2$	x_1~Ber(0.5), x_2~U(-3,3)	**3.6**	**3.5**	**3.7**	**3.5**	4.7	4.7
8	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3$	x_1, x_2, x_3~U(-6,6)	**4.5**	4.8	4.7	4.9	5.0	4.8
9	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3$	x_1, x_2, x_3~N0,1.5)	4.8	5.2	4.9	5.2	5.0	5.2
10	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3$	x_1, x_2, x_3~U(-6,6), N(0,1.5), $\chi^2(4)$	4.6	4.6	4.7	4.7	5.1	4.9
11	$\beta_0+\beta_1 x_1+\beta_2 x_1^2$	x_1~U(-3,3) W=0.01	**1.8**	**2.9**	**2.0**	**2.9**	**3.6**	4.7
12	$\beta_0+\beta_1 x_1+\beta_2 x_1^2$	x_1~U(-3,3) W=0.05	**2.6**	**3.3**	**2.6**	**3.4**	**4.1**	5.1
13	$\beta_0+\beta_1 x_1+\beta_2 x_1^2$	x_1~U(-3,3) W=0.1	**3.5**	**3.8**	**3.6**	**4.0**	**4.5**	4.7
14	$\beta_0+\beta_1 x_1+\beta_2 x_1^2$	x_1~U(-3,3) W=0.2	**4.2**	**4.2**	**4.5**	**4.2**	5.1	4.8
15	$\beta_0+\beta_1 x_1+\beta_2 x_1^2$	x_1~U(-3,3) W=0.4	**4.4**	**4.3**	**4.5**	**4.4**	4.9	4.9
16	$\beta_0+\beta_1 x_1+\beta_2 x_1^2+x_2$	x_1, x_2~U(-3,3) W=0.01	**5.5**	**5.6**	**5.7**	**5.7**	**6.1**	**5.9**
17	$\beta_0+\beta_1 x_1+\beta_2 x_1^2+x_2$	x_1, x_2~U(-3,3) W=0.05	**6.5**	5.3	**6.7**	5.4	**7.1**	**5.8**
18	$\beta_0+\beta_1 x_1+\beta_2 x_1^2+x_2$	x_1, x_2~U(-3,3) W=0.1	**6.0**	5.2	**6.2**	5.3	**6.5**	5.4
19	$\beta_0+\beta_1 x_1+\beta_2 x_1^2+x_2$	x_1, x_2~U(-3,3) W=0.2	**6.3**	4.9	**6.5**	5.1	**6.9**	5.3
20	$\beta_0+\beta_1 x_1+\beta_2 x_1^2+x_2$	x_1, x_2~U(-3,3) W=0.4	**6.4**	**5.6**	**6.5**	**5.8**	**6.9**	**5.7**

Table 4.4 Simulated null rejection per cent† (n=100 and 500, r=10,000, α=0.05) for settings 1-24. (cont.)

Setting	Linear predictor	Covariate distributions§	*HL*		J^2*		*T*	
			100	500	100	500	100	500
21	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	x_1~U(-3,3), x_2=(0,1) V=0.01	**2.2**	**3.1**	**2.2**	**3.1**	**4.4**	4.9
22	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	x_1~U(-3,3), x_2=(0,1) V=0.03	**2.6**	**2.6**	**2.6**	**2.6**	4.9	5.0
23	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	x_1~U(-3,3), x_2=(0,1) V=0.05	**2.4**	**2.7**	**2.5**	**2.7**	5.0	5.2
24	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	x_1~U(-3,3), x_2=(0,1) V=0.07	**2.2**	**2.3**	**2.2**	**2.4**	4.8	5.0

* The distribution of J^2 was assumed to be $\chi^2(8)$.

§ For settings 11-20, the coefficients were chosen such that the model curve passed through the points $\pi(-1.5)=0.05$, $\pi(3)=0.95$, and $\pi(-3)=W$.

For settings 21-24, the coefficients were chosen such that the model curve passed through the points $\pi(-3,0)=0.1$, $\pi(-3,1)=0.1$, $\pi(3,0)=0.2$, and $\pi(3,1)=0.2 + V$.

† Rejection percentages were significantly different from 5% at α=0.05 if they fell above 5.43 or below 4.57 (bold).

$n = 500$) than HL and J^2 (25% for $n = 100$, 46% for $n = 500$ in both cases). This was reflected in the overall rejection percentages of T, which were 5.3 $(n = 100)$ and 5.1 $(n = 500)$. In contrast, those of HL and J^2, were 4.4 and 4.5 for $n = 100$, and 4.2 and 4.3 for $n = 500$, respectively. We note that the results for HL in Table 4.3 are similar to those reported in Hosmer, et al. (1997) . Among the 24 null settings considered, the test statistics agreed on whether to reject the null hypothesis 97 to 98% of the time. When the statistics did disagree, most of the time HL and J^2 agreed with each other and disagreed with T.

4.5.2.3 Power - Rejection Percentage Under the Alternative Hypothesis

Simulations investigating the power of HL, J^2, and T to detect different types of departure from a true underlying model were performed. Our results for HL agreed closely with those reported by Xie (2005) and Xie, et al. (2008). Our results also agreed with those reported by Hosmer, et al. (1997), except for the incorrectly specified link simulations, for which some difference between results was observed.

The power of each statistic to detect the omission of a quadratic term from the fitted model was examined, and the rejection percentages are presented in Table 4.5. Each of the statistics had low ($< 33\%$) power to detect a slight departure from linearity (setting 11 and setting 12 ($n = 100$ only)), but the power quickly increased to moderate to high levels ($51 - 100\%$) as the influence of the quadratic term increased (settings 13-15). As might be expected, this increase in power occurred more rapidly when the sample size was larger. When these same models included an additional covariate (settings 16-20), all statistics had low power to detect the omitted quadratic term. Among these 10 settings and two sample numbers of observations, the test statistics agreed on whether to reject the null hypothesis 95 to 100% of the time. When the statistics did disagree, most of the time HL and J^2 agreed with each other and disagreed with T.

All of the statistics had low $(\leq 7\%)$ power to detect the omission of dichotomous and interaction terms in the settings studied (settings 21-24), regardless of sample size. The

rejection percentages for these simulations are also given in Table 4.5. Among these 5 settings and two levels of observations, the test statistics agreed on whether to reject the null hypothesis 97 to 98% of the time. When the statistics did disagree, most of the time HL and J^2 agreed with each other and disagreed with T.

The power of HL to detect an alternative link (settings 25-29) was variable, with the rejection per cent ranging from 4.2 to 92.4 for the settings considered. Our results are reported in Table 4.6. When $n = 100$, the power to detect the use of an incorrect link was very low $(\leq 6\%)$ except when the underlying model had an asymmetric tails link. In this case it was slightly higher $(\sim 15\%)$. For $n = 500$, the results were more variable. All of the statistics had low power $(\leq 8\%)$ to detect an incorrectly specified logistic model when the underlying model had a probit or long-tail link. The power to detect departure was slightly greater when the underlying model had a short tails or complementary log-log link $(17 - 28\%)$, and it was very strong $(> 91\%)$ when the model had an asymmetric tails link. Among these 5 settings and two sample sizes, the test statistics agreed on whether to reject the null hypothesis 93 to 99% of the time. When the statistics did disagree, most of the time HL and J^2 agreed with each other and disagreed with T.

Table 4.5 Power of *HL*, J^2 and *T* to detect a logistic model with an incorrectly specified linear predictor.

	True	Specified	*HL*		J^2*		*T*	
Setting	Linear Predictor	Linear Predictor	100	500	100	500	100	500
11	$\beta_0+\beta_1x_1+\beta_2x_1^2$	$\beta_0+\beta_1x_1$	7.4	7.5	7.6	7.6	7.3	7.5
12	$\beta_0+\beta_1x_1+\beta_2x_1^2$	$\beta_0+\beta_1x_1$	32.0	80.1	32.2	80.3	30.8	78.6
13	$\beta_0+\beta_1x_1+\beta_2x_1^2$	$\beta_0+\beta_1x_1$	53.0	98.3	53.2	98.3	51.4	98.1
14	$\beta_0+\beta_1x_1+\beta_2x_1^2$	$\beta_0+\beta_1x_1$	76.2	100	76.3	100	74.5	100
15	$\beta_0+\beta_1x_1+\beta_2x_1^2$	$\beta_0+\beta_1x_1$	93.8	100	93.8	100	92.7	100
16	$\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3$	$\beta_0+\beta_1x_1+\beta_3x_3$	6.3	6.3	6.5	6.4	6.4	6.4
17	$\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3$	$\beta_0+\beta_1x_1+\beta_3x_3$	8.4	10.7	8.5	10.9	8.5	10.5
18	$\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3$	$\beta_0+\beta_1x_1+\beta_3x_3$	10.0	13.1	10.2	13.4	10.1	12.7
19	$\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3$	$\beta_0+\beta_1x_1+\beta_3x_3$	10.6	17.1	10.9	17.5	10.8	16.8
20	$\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3$	$\beta_0+\beta_1x_1+\beta_3x_3$	10.5	19.3	10.7	19.9	10.4	19.5
21	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	$\beta_0+\beta_1x_1$	4.0	4.8	4.0	4.8	4.0	4.7
22	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	$\beta_0+\beta_1x_1$	3.9	5.2	4.0	5.2	4.1	4.9
23	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	$\beta_0+\beta_1x_1$	4.6	5.0	4.7	5.1	4.6	5.1
24	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	$\beta_0+\beta_1x_1$	4.6	6.8	4.7	6.9	4.5	7.0

* The distribution of J^2 is assumed to be $\chi^2(8)$.

Table 4.6 Power of *HL*, J^2 and *T* to detect a Stukel generalized model with an incorrectly specified logistic link function.

			HL		J^2*		*T*	
Setting	True Model	Fitted Model	100	500	100	500	100	500
25	Probit	Logit	4.3	4.9	4.4	5.1	4.2	5.0
26	Long tail	Logit	5.3	7.5	5.3	7.6	5.3	8.0
27	Short tail	Logit	2.3	17.6	2.5	18.3	2.6	17.2
28	Complementary log-log	Logit	5.8	27.2	6.0	27.6	5.5	26.1
29	Asymptotic long-short	Logit	14.5	92.3	14.7	92.4	14.1	91.1

* The distribution of J^2 is assumed to be $\chi^2(8)$.

4.6 Examples

Our simulations indicate that if HL and J^2 are assumed to follow a $\chi^2(G-2)$ distribution and T a $\chi^2(G-1)$ distribution, their assessments of model fit are similar. However, we showed in section 4.4 that the difference between HL and J^2 may be large if any group contains consecutive predicted probabilities whose difference is close to 1. A difference between two consecutive predicted probabilities this large was not observed in the simulations studied. Thus, to illustrate this we considered two examples with real data sets from the Low Birth Weight study described in *Applied Logistic Regression* (2000) and one example with a computer generated data set. In both of the real data examples, a binary logistic regression model was chosen to predict the incidence of low birth weights among babies, using covariates that relate to their mother's behaviour and physical characteristics. The first model considered is discussed in chapter 2 of *Applied Logistic Regression* (2000), and contains race (white, black, other) and the weight of the mother at her last menstrual period as covariates. The second model uses the mother's age, her smoking status, and the presence of uterine irritability as covariates. Both have logits that are linear in their variables. The values of HL, J^2, and T were calculated using grouping criteria (e.g. how to deal with ties) from a standard commercial software package (StataCorp 2007). In the case of the first model, each of the statistics indicated that the model is a fairly good fit to the data $\left(HL = 7.60 \ \left(p = 0.47\right),\ J^2 = 7.61 \ \left(p = 0.47\right), \text{ and } T = 8.20 \ \left(p = 0.51\right)\right)$. In the second example, each of the statistics indicate a poor fit to the data $\left(HL = 15.84 \ \left(p = 0.045\right),\right.$ $\left.J^2 = 15.86 \ \left(p = 0.044\right), \text{ and } T = 16.78 \ \left(p = 0.052\right)\right)$. In the first case the largest difference between any two consecutive predicted probabilities within one group was 0.12, while in the second model the largest difference was 0.18. In both cases, HL and J^2 are very close and the probability values (p) of the three statistics are very similar.

Our third example was simulated to mimic a real world example using one continuous covariate and one dichotomous covariate. Five hundred covariate values of x_1 were

randomly sampled (using Stata 10.1 with seed 13988101) from a $N(0.7, 0.2)$ distribution, and then 1.6 was added to 275 observations, making the distribution of x_1 bimodal with one of the decile groups containing values from each mode. The first 225 observations are distinguished from the last 275 observations using an indicator variable D. Probabilities were then generated using a logit link with linear predictor, $-5 + 2x_1 + 0.7x_1^2$. Binary outcomes of y were simulated by comparing the generated probabilities to a value u where $u \sim U(0,1)$ according to the rule $y = I(u < \pi)$, where I is an indicator function such that $I = 1$ when the argument is true, and $I = 0$ otherwise. An analyst would be given 3 fields, the outcome, an indicator variable D that could represent gender and the continuous variable x_1 that would represent some biological output. We used Stata's frapoly command (StataCorp 2007) to perform a fractional polynomial regression. We found that x_1 is linear in the logit of y and that there is no interaction between x_1 and D. That is, the outcome is best predicted in terms of x_1 and D. Hence, in truth the final model is incorrectly specified. The calculated statistics from this example were $HL = 12.75$ $(p = 0.121)$, $J^2 = 15.95$ $(p = 0.043)$, and $T = 16.00$ $(p = 0.067)$. In this case J^2 correctly rejected the incorrectly specified model, but HL and T did not. The same simulation was run again under the null, using the covariates x_1 and x_1^2, and the seed 12268179. The calculated results were $HL = 13.55$ $(p = 0.094)$, $J^2 = 15.624$ $(p = 0.045)$, and $T = 15.850$ $(p = 0.070)$. In this case J^2 incorrectly rejected the correctly specified model, but HL and T did not. Finally, the null scenario was rerun again (seed 20859170). Here the calculated values of the statistics were $HL = 11.793$ $(p = 0.161)$, $J^2 = 12.429$ $(p = 0.133)$, and $T = 22.594$ $(p = 0.007)$, giving an example where HL and J^2 correctly fail to reject the null hypothesis while T incorrectly rejects it. In the three bimodal examples the greatest difference between predicted probabilities within one group was 0.8. This was much larger than the differences observed in the low birth weight study, causing the difference between

the values of HL and J^2 in the bimodal cases to be greater. Neither rejected the null hypothesis, however.

4.7 Discussion

We have shown that when the deciles-of-risk grouping method is used, the HL, J^2 and T goodness-of-fit statistics are closely related algebraically. In fact, the tests are equivalent when all covariate patterns within each of the groups are the same. It has been shown that random cells created when the grouping rule is based on estimated parameters reduces the degrees of freedom of the distribution of HL (Hosmer, et al. 1980) from $\chi^2(G-1)$ in the case when the outcomes are multinomial and the cells are not random, to approximately $\chi^2(G-2)$ when at least one covariates is continuous and the boundaries are random. Halteman (1980) showed that the asymptotic distribution of T is not affected by random cells. The effect of random cells on J^2 has not, to our knowledge, been reported in the literature, although Pigeon, et al. (1999b) state that the choice of grouping method does not affect its distribution. Our results suggest that this is not the case, at least in the scenarios we examined. Our simulations indicate that when the deciles-of-risk grouping method is used, the null distributions reported for HL and T are indeed those reported (i.e. approximately $\chi^2(G-2)$ for HL (Hosmer, et al. 1980) and approximately $\chi^2(G-1)$ for T (Halteman 1980)). Our simulations gave stronger evidence in the case of T than the simulations reported by Halteman (1980). We however found the null distribution of J^2, when $G=10$, to be $\chi^2(8)$ rather than the $\chi^2(9)$ reported by Pigeon, et al. (1999b). Further research on how random cells affect the distribution of J^2 is needed.

Our simulations with large numbers of replication showed that the distributions of the three test statistics were most similar to their respective chi-squared distributions when the linear predictor contained a covariate with a uniform distribution rather than with a skewed

distribution, such as a chi-squared distribution. This suggests that the distributions of all three of the statistics are affected by the distribution of the continuous covariates in the model.

When HL and J^2 were assumed to follow a $\chi^2(G-2)$ distribution and T a $\chi^2(G-1)$ distribution, the empirical Type I error rate was controlled by T about twice as often as HL or J^2. All three statistics had similar power to detect incorrectly specified models, and agreed on whether to reject the null hypothesis most of the time $(>97\%)$. When there was disagreement among the tests, usually decisions based on T disagreed with those based on the other two statistics. This is interesting since, as a consequence of Theorem 4.1 and Theorem 4.2, the rejection rate under the null hypothesis and the power under the alternative hypothesis must always be greater for J^2 than for HL. However, the close agreement between HL and J^2 is not entirely surprising because we did not specifically look at cases where their difference would be very large. That is, our simulations did not specifically look at cases where one of the decile groups contained two consecutive predicted probabilities whose difference was near 1. In our simulated example however, when a difference between two predicted probabilities is much closer to 1, we demonstrated that examples exist where J^2 is much greater than HL, and where T is much greater than J^2.

We investigated whether these types of cases would produce any difference in the distributional properties and the performances of the three statistics (data not shown). We reran the analysis introduced in section 6 using the settings in the third example for $n=500$. We found that the null rejection percentages amongst the three statistics were similar, all controlled the Type I error rate, and all had low power $(<10\%)$ to detect an incorrectly specified model, although there was a significant difference $(p=0.004)$ between the power of HL and T, with T having the greater power in our simulations (7.8 vs. 9.6 per cent

rejection). Although there were some differences between the three statistics, in general we found that the statistics all performed similarly under this bimodal setting.

Hosmer, et al. (1997) note that when a goodness-of-fit test uses a grouping method that pre-specifies the group boundaries, decisions on model fit may be more influenced by the choice of boundary cut-points rather than by the lack of fit of the model. Because the deciles-of-risk grouping method is more standardized than partitioning the covariate space and more easily implemented, it can be a more attractive option. If the deciles-of-risk method is chosen as a grouping method, our results show that T controlled the Type I error rate about twice as often as HL and J^2, but all three had similar power to detect an incorrectly specified model. Although HL and J^2 performed similarly, more work needs to be done to clarify the distributional properties of J^2. The properties of HL under the null distribution using the deciles-of-risk partitioning method have been extensively studied, and this test is easily implemented and understood by most users. Because of this, among these test statistics we recommend either HL or T for validating the goodness-of-fit of logistic models when the deciles-of-risk grouping method is used.

Chapter 5 Proposed Goodness-of-Fit Statistic for Binary GLM with Non-Canonical Links

5.1 Expanded Tsiatis model

Several GLMs with different link functions have been used to model binary data. These include the logit, probit, log-log, complementary log-log, and log binomial models (Hardin, et al. 2007). The model and link functions, as well as the canonical parameter evaluated under these models, are described in section 2.6.

The original Tsiatis goodness-of-fit score test (Tsiatis 1980) was developed to assess the fit of logistic regression models to observed outcome data. Under the canonical logit link, the elements of $\mathbf{S}$, (3.12), can be expressed using the chain rule (Hardin, et al. 2007) as

$$\frac{\partial l}{\partial \gamma_g} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \zeta_i} \frac{\partial \zeta_i}{\partial \pi(\mathbf{x}_i, \mathbf{I}_i)} \frac{\partial \pi(\mathbf{x}_i, \mathbf{I}_i)}{\partial \eta_i} \frac{\partial \eta_i}{\partial \gamma_g} \tag{5.1}$$

$$= \sum_{i=1}^{n} \left\{ y_i - \frac{\partial b(\zeta)_i}{d\zeta_i} \right\} \frac{\partial \zeta_i}{\partial \eta_i} \frac{\partial \left\{ \mathbf{x}_i'\boldsymbol{\beta} + \sum_{g=1}^{G} \gamma_g I_i^{(g)} \right\}}{\partial \gamma_g} \tag{5.2}$$

$$= \sum_{i=1}^{n} \left\{ y_i - \pi(\mathbf{x}_i, \mathbf{I}_i) \right\} I_i^{(g)} \tag{5.3}$$

because $\zeta = \eta$ under the canonical link. Here $\left\{ I_i^{(1)}, ..., I_i^{(G)} \right\}$ are a set of indicator functions for the ith observation that are defined as $I_i^{(g)} = 1$ when the covariates lie in it the gth region, and $I_i^{(g)} = 0$ otherwise, and $\{\gamma_1, ..., \gamma_G\}$ is the set of additional coefficients associated with each of the G indicator functions.

To our knowledge, the application of the Tsiatis score test to non-canonical GLM for binary outcomes has not been presented in the literature. A problem with making this direct application is that under non-canonical link functions, the canonical parameter, ζ, and the linear predictor, η, are not equal, and thus the second term on the right-hand side of (5.2) does not cancel, and (5.3) does not result in the "observed minus expected" form that is

desired for a goodness-of-fit test statistic. Instead, under any non-canonical link function, the resulting expression is

$$\frac{\partial l}{\partial \gamma_g} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \zeta_i} \frac{\partial \zeta_i}{\partial \theta(\mathbf{x}_i, \mathbf{I}_i)} \frac{\partial \theta(\mathbf{x}_i, \mathbf{I}_i)}{\partial \eta_i} \frac{\partial \eta_i}{\partial \gamma_g} \tag{5.4}$$

$$= \sum_{i=1}^{n} \frac{\partial l_i}{\partial \zeta_i} \frac{\partial \zeta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \gamma_g} \tag{5.5}$$

$$= \sum_{i=1}^{n} \{ y_i - \theta(\mathbf{x}_i, \mathbf{I}_i) \} \frac{\partial \zeta_i}{\partial \eta_i} I_i^{(g)} \tag{5.6}$$

Here θ is used in place of π to differentiate between the probabilities under the canonical and non-canonical links.

A new model is proposed that is an augmented version of the usual GLM model for Bernoulli outcomes,

$$\theta(\mathbf{x}) = g^{-1}(\mathbf{x}'\boldsymbol{\beta}) \tag{5.7}$$

and is an analog of the Tsiatis model (3.9) (Tsiatis 1980). Here (5.7) is referred to as the null model, and has the linear predictor

$$\eta_0 = \mathbf{x}'\boldsymbol{\beta} \tag{5.8}$$

Under the null model the canonical parameter is expressed as

$$\zeta_0 = \ln\left(\frac{\theta(\eta_o)}{1 - \theta(\eta_o)} \right) \tag{5.9}$$

The augmented model has terms that are added to η_0 so that, under any link function, the resulting test statistic will be a quadratic form of observed minus expected count, as is the case with the Tsiatis statistic under the canonical logit link, as it is in (5.3). This augmented model allows for the assessment of the fit of the null model via a score test. The model is referred to here as the *generalized Tsiatis model* and is expressed as

$$\theta(\mathbf{x}, \mathbf{I}) = g^{-1}(\eta) \tag{5.10}$$

where the linear predictor, η, of the new model is

$$\eta = \mathbf{x}'\boldsymbol{\beta} + h(\mathbf{x}'\boldsymbol{\beta})\sum_{g=1}^{G}\gamma_g I^{(g)} \tag{5.11}$$

As in the case of the original Tsiatis model, $\left\{I^{(1)},...,I^{(G)}\right\}$ is a set of indicator functions defined as $I^{(g)}=1$ when the *i*th observation is in the *g*th region (or group under the deciles-of-risk method), and $I^{(g)}=0$ otherwise, and $\{\gamma_1,...,\gamma_G\}$ are the set of constant terms that are multiplied by the respective indicator function. The $\sum_{g=1}^{G}\gamma_g I^{(g)}$ term is multiplied by an additional term that is the inverse of the second and third terms of (5.4) under the null model. That is,

$$h(\mathbf{x}_i'\boldsymbol{\beta}) = \left[\frac{\partial\zeta_{i0}}{\partial\hat{\theta}(\mathbf{x}_i)}\frac{\partial\hat{\theta}(\mathbf{x}_i)}{\partial\eta_{i0}}\right]^{-1} = \frac{\partial\eta_{i0}}{\partial\zeta_{i0}} \tag{5.11}$$

The term $h(\mathbf{x}'\boldsymbol{\beta})$ has the effect of transforming the form of (5.6) into the desired "observed minus expected" form of (5.3) for any Bernoulli GLM model. Note that when the link function is canonical , then $\zeta_0=\eta_0$ and $h(\mathbf{x}'\boldsymbol{\beta})=1$, and $h(\mathbf{x}_i'\boldsymbol{\beta})\sum_{g=1}^{G}\gamma_g I_i^{(g)} = \sum_{g=1}^{G}\gamma_g I_i^{(g)}$. Thus, in the canonical case, the augmentation of the model by $h(\mathbf{x}_i'\boldsymbol{\beta})\sum_{g=1}^{G}\gamma_g I_i^{(g)}$ will be constant over the entire *g*th region or group. When the link is non-canonical, however, $h(\mathbf{x}_i'\boldsymbol{\beta})\sum_{g=1}^{G}\gamma_g I_i^{(g)}$ varies as a function of the covariate data.

5.2 Generalized Tsiatis GOF Statistic

A new test statistic, $T_{\mathcal{G}}$, is proposed to assess the goodness-of-fit of GLMs for binary data. It is an adaptation of the original Tsiatis score test statistic for logistic regression models (Tsiatis 1980) that can be used to evaluate a GLM with a non-canonical link function. The

null hypothesis tested is that $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_G) = \mathbf{0}$, and thus under the null hypothesis (5.7) and (5.10) are equal.

The log-likelihood of the new expanded model is

$$l = \log L = \sum_{i=1}^{n} y_i \ln \theta(\mathbf{x}_i, \mathbf{I}_i) + (1 - y_i) \ln \{1 - \theta(\mathbf{x}_i, \mathbf{I}_i)\} \tag{5.12}$$

The beta coefficients of the model are treated as nuisance parameters with respect to assessing goodness-of-fit, and are estimated using maximum likelihood estimation under the null hypothesis and expressed as $\hat{\boldsymbol{\beta}}$. The goodness-of-fit score test statistic, conditional on $\hat{\boldsymbol{\beta}}$, is calculated as

$$T_{\mathcal{G}} = \mathbf{S}'\mathbf{V}^{-1}\mathbf{S} \tag{5.13}$$

where $\mathbf{S}$ is the $G \times 1$ score vector $(\partial l/\partial \gamma_1, ..., \partial l/\partial \gamma_G)'$, and $\mathbf{V}$ is the $G \times G$ conditional covariance matrix

$$\mathbf{V} = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}' \tag{5.14}$$

with the general entries of the matrices of (5.14) being

$$A_{gg'} = \begin{cases} -E\left(\dfrac{\partial^2 l}{\partial \gamma_g \partial \gamma_{g'}}\right) & (g = g';\ g, g' = 1, ..., G) \\ 0 & (g \neq g') \end{cases} \tag{5.15}$$

$$B_{gk} = -E\left(\frac{\partial^2 l}{\partial \gamma_g \partial \beta_k}\right) \qquad (g = 1, ..., G,\ \ k = 0, ..., K) \tag{5.16}$$

$$C_{kk'} = -E\left(\frac{\partial^2 l}{\partial \beta_k \partial \beta_{k'}}\right)\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=\mathbf{0}} \qquad (k, k' = 0, ..., K) \tag{5.17}$$

All of the terms of $T_{\mathcal{G}}$ are evaluated under $\boldsymbol{\gamma} = \mathbf{0}$ and $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimates of $\boldsymbol{\beta}$ in the null model (5.7). Under the null hypothesis, $\theta(\mathbf{x}, \mathbf{I}) = \theta(\mathbf{x})$, and so in general will be designated as $\hat{\theta}$.

The gth element of the score vector $\mathbf{S}$ is expressed as

$$\frac{\partial l}{\partial \gamma_g} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \zeta_i} \frac{\partial \zeta_i}{\partial \theta(\mathbf{x}_i, \mathbf{I}_i)} \frac{\partial \theta(\mathbf{x}_i, \mathbf{I}_i)}{\partial \eta_i} \frac{\partial \eta_i}{\partial \gamma_g} \qquad (g = 1, ..., G) \tag{5.18}$$

$$= \sum\nolimits_{i=1}^{n} \frac{\partial l_i}{\partial \zeta_i} \frac{\partial \zeta_i}{\partial \theta(\mathbf{x}_i, \mathbf{I}_i)} \frac{\partial \theta(\mathbf{x}_i, \mathbf{I}_i)}{\partial \eta_i} \frac{\partial \left\{ \mathbf{x}_i'\boldsymbol{\beta} + h(\mathbf{x}_i'\boldsymbol{\beta}) \sum_{g=1}^{G} \gamma_g I_i^{(g)} \right\}_i}{\partial \gamma_g} \tag{5.19}$$

$$= \sum\nolimits_{i=1}^{n} \{ y_i - \theta(\mathbf{x}_i, \mathbf{I}_i) \} \frac{\partial \zeta_i}{\partial \eta_i} \frac{\partial \eta_{i0}}{\partial \zeta_{i0}} I_i^{(g)} \tag{5.20}$$

which under the null hypothesis evaluates to

$$= \sum_{i=1}^{n} \left(y_i - \hat{\theta} \right) I_i^{(g)} \tag{5.21}$$

Using (2.9), (2.10), and (2.14), the elements in $\mathbf{V}$ are

$$A_{gg'} = \frac{1}{\operatorname{var}\{\theta(\mathbf{x}_i, \mathbf{I}_i)\}} \left(\frac{\partial \theta(\mathbf{x}_i, \mathbf{I}_i)}{\partial \eta} \right)^2 \frac{\partial \eta}{\partial \gamma_g} \frac{\partial \eta}{\partial \gamma_{g'}} \tag{5.22}$$

$$= \left(\frac{\partial \theta(\mathbf{x}, \mathbf{I})}{\partial \eta} \right)_i^2 \left(\frac{\partial \zeta}{\partial \theta(\mathbf{x}, \mathbf{I})} \right)_i \left(\frac{\partial \eta_0}{\partial \zeta_0} \right)_i^2 I_i^{(g)} I_i^{(g')} \tag{5.23}$$

which under the null hypothesis becomes

$$A_{gg'}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \boldsymbol{\gamma}=0} = \left(\frac{\partial \hat{\theta}}{\partial \eta_0} \right)_i^2 \left(\frac{\partial \zeta_0}{\partial \hat{\theta}} \right)_i \left(\frac{\partial \eta_0}{\partial \zeta_0} \right)_i^2 I_i^{(g)} I_i^{(g')} \tag{5.24}$$

$$= \hat{\theta}\left(1 - \hat{\theta}\right) I_i^{(g)} I_i^{(g')} \tag{5.25}$$

and

$$B_{gk} = \sum\nolimits_{i=1}^{n} \frac{1}{\operatorname{var}\{\theta(\mathbf{x}_i, \mathbf{I}_i)\}} \left(\frac{\partial \theta(\mathbf{x}_i, \mathbf{I}_i)}{\partial \eta} \right)^2 \frac{\partial \eta}{\partial \gamma_g} \frac{\partial \eta}{\partial \beta_k} \tag{5.26}$$

$$= \sum_{i=1}^{n} \frac{1}{\theta(\mathbf{x}_i, \mathbf{I}_i)\{1-\theta(\mathbf{x}_i, \mathbf{I}_i)\}} \left(\frac{\partial \theta(\mathbf{x}_i, \mathbf{I}_i)}{\partial \eta} \right)^2 \{h(\mathbf{x}_i'\boldsymbol{\beta})\} I_i^{(g)} \left[x_{ij} + \frac{\partial h(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \beta_k} \sum_{g=1}^{G} \gamma_g I_i^{(g)} \right] \tag{5.27}$$

which under the null hypothesis becomes

$$B_{gk}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \boldsymbol{\gamma}=0} = \sum_{i=1}^{n} \left(\frac{\partial \hat{\theta}}{\partial \eta_0} \right)_i^2 \left(\frac{\partial \zeta_0}{\partial \hat{\theta}} \right)_{0i} \left(\frac{\partial \eta_0}{\partial \zeta_0} \right)_i I_i^{(g)} x_{ik} \tag{5.27}$$

$$= \sum_{i=1}^{n} \left(\frac{\partial \hat{\theta}}{\partial \eta_0} \right)_i^2 I_i^{(g)} x_{ik} \tag{5.27}$$

and

$$C_{kk'} = \frac{1}{\operatorname{var}\{\theta(\mathbf{x}_i, \mathbf{I}_i)\}} \left(\frac{\partial \theta(\mathbf{x}_i, \mathbf{I}_i)}{\partial \eta} \right)^2 \frac{\partial \eta}{\partial \beta_k} \frac{\partial \eta}{\partial \beta_{k'}} \tag{5.28}$$

$$= \sum_{i=1}^{n} \left(\frac{\partial \zeta}{\partial \theta(\mathbf{x}, \mathbf{I})} \right)_i \left(\frac{\partial \theta(\mathbf{x}, \mathbf{I})}{\partial \eta} \right)_i^2 \left[x_{ik} + \frac{\partial h(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \beta_k} \sum_{g=1}^{G} \gamma_g I_i^{(g)} \right] \left[x_{ik'} + \frac{\partial h(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \beta_{k'}} \sum_{g=1}^{G} \gamma_g I_i^{(g)} \right] \tag{5.29}$$

which under the null hypothesis is

$$C_{kk'}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \boldsymbol{\gamma}=0} = \sum_{i=1}^{n} \left(\frac{\partial \hat{\theta}}{\partial \eta_0} \right)_i^2 \left(\frac{\partial \zeta_0}{\partial \hat{\theta}} \right)_i x_{ik} x_{ik'} \tag{5.30}$$

5.3 Forms of $T_{\mathcal{G}}$ Under Several Common Link Functions

The $T_{\mathcal{G}}$ statistic can be adapted to assess the fit of the logit, probit, log-log, complementary log-log, and log binomial models. The symbol $T_{\mathcal{G}}$ will be used when referring to the generalized Tsiatis statistics calculated under any link function, or to a group of these statistics. However, under the non-canonical, probit, log-log, complementary log-log, and log links, the $T_{\mathcal{G}}$ statistic will be designated as T_{Pr}, T_{LL}, T_{Cll}, and T_{LB}, respectively. Under

the canonical logit link $T_{\mathcal{G}}$ is the original T statistic. In general, the predicted probability under any link will be designated as $\hat{\theta}_{\mathcal{G}}$. Under the non-canonical probit, log-log, complementary log-log, and log links, the predicted probabilities will be represented by $\hat{\theta}_{\text{Pr}}$, $\hat{\theta}_{LL}$, $\hat{\theta}_{Cll}$, and $\hat{\theta}_{LB}$ respectively. The general form of the terms in the score vector are (5.21), and are designated in general as

$$\left.\frac{\partial l}{\partial \gamma_g}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum_{i=1}^{n}\left(y_i - \hat{\theta}_{\mathcal{G}}\right) I_i^{(g)} \tag{5.31}$$

where $\hat{\theta}_{\mathcal{G}}$ may be replaced by any of the links studied. The terms necessary for the evaluation of $\mathbf{V}$, that is terms (5.25), (5.27), and (5.30) , evaluated under the probit, log-log, complementary log-log, and log links are given in Table 5.1, and their derivations are presented in Appendix A.

5.4 Distribution and Degrees of Freedom of $T_{\mathcal{G}}$

Score test statistics have distributions that are asymptotically χ^2 with degrees of freedom equal to the dimension of the vector of parameters of interest, which is also the rank of the covariance matrix (Kendall, et al. 1999, Smyth 2003). In his dissertation, Halteman (1980) derived the same GOF score test statistic for assessing the fit of logistic regression models as the one reported by Tsiatis (1980) , T . He presented a theorem that states that the sum of any row or any column of $\mathbf{V}$ is zero, and thus that the rank of $\mathbf{V}$ can be at most $G-1$, and so is always singular. It is easy to show a counter example to prove that this does not always hold for $T_{\mathcal{G}}$ when evaluating non-canonical models. In fact, in our simulations described below in section 5.8.1.2, which are used to study the distribution of $T_{\mathcal{G}}$ for the non-canonical links, every replicate (1.2 million) returned a $\mathbf{V}$ matrix with a rank of $G=10$. Thus by counter example, Theorem 5.1 does not hold for the non-canonical link functions we studied, and thus it is possible for $\mathbf{V}$ to be nonsingular. This is possible because under the non-canonical links, the canonical link function ζ and the linear predictor η are not

Table 5.1 General elements of the covariance matrix **V**, used in the calculation of $T_{G,}$ under the logit, probit, log-log, complementary log-log, and log links.

Link	$A_{gg'}\big\|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\,\boldsymbol{\gamma}=0}$	$B_{gk}\big\|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\,\boldsymbol{\gamma}=0}$	$C_{kk'}\big\|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\,\boldsymbol{\gamma}=0}$
Logit	$\sum_{i=1}^{n}\hat{\pi}_i\left(1-\hat{\pi}_i\right)I_i^{(g)}I_i^{(g')}$	$\sum_{i=1}^{n}\hat{\pi}_i\left(1-\hat{\pi}_i\right)I_i^{(g)}x_{ik}$	$\sum_{i=1}^{n}\hat{\pi}_i\left(1-\hat{\pi}_i\right)x_{ik}x_{ik'}$
Probit	$\sum_{i=1}^{n}\hat{\theta}_{\mathrm{Pr}_i}\left(1-\hat{\theta}_{\mathrm{Pr}_i}\right)I_i^{(g)}I_i^{(g')}$	$\sum_{i=1}^{n}\hat{\phi}I_i^{(g)}x_{ik}$	$\sum_{i=1}^{n}\frac{\hat{\phi}^2}{\hat{\theta}_{\mathrm{Pr}_i}\left(1-\hat{\theta}_{\mathrm{Pr}_i}\right)}x_{ik}x_{ik'}$
Log-log	$\sum_{i=1}^{n}\hat{\theta}_{LL_i}\left(1-\hat{\theta}_{LL_i}\right)I_i^{(g)}I_i^{(g')}$	$\sum_{i=1}^{n}-\hat{\theta}_{LL_i}\ln\hat{\theta}_{LL_i}I_i^{(g)}x_{ik}$	$\sum_{i=1}^{n}\frac{\hat{\theta}_{LL_i}}{1-\hat{\theta}_{LL_i}}\left(\ln\hat{\theta}_{LL_i}\right)^2x_{ik}x_{ik'}$
Complementary Log-log	$\sum_{i=1}^{n}\hat{\theta}_{Cll_i}\left(1-\hat{\theta}_{Cll_i}\right)I_i^{(g)}I_i^{(g')}$	$\sum_{i=1}^{n}-\left(1-\hat{\theta}_{Cll_i}\right)\ln\left(1-\hat{\theta}_{Cll_i}\right)I_i^{(g)}x_{ik}$	$\sum_{i=1}^{n}\frac{1-\hat{\theta}_{Cll_i}}{\hat{\theta}_{Cll_i}}\left\{\ln\left(1-\hat{\theta}_{Cll_i}\right)\right\}^2x_{ik}x_{ik'}$
Log	$\sum_{i=1}^{n}\hat{\theta}_{LB_i}\left(1-\hat{\theta}_{LB_i}\right)I_i^{(g)}I_i^{(g')}$	$\sum_{i=1}^{n}\hat{\theta}_{LB_i}I_i^{(g)}x_{ik}$	$\sum_{i=1}^{n}\frac{\hat{\theta}_{LB_i}}{1-\hat{\theta}_{LB_i}}x_{ik}x_{ik'}$

equal as they are under the canonical logit link. To see in detail why this occurs, we present Halteman's theorem using our notation.

Halteman begins by algebraically manipulating the expression for the sum across the gth row of the matrix, $\mathbf{V}$, (3.13). He then shows, using an expression due to (Rao 1973) for the inverse of $\mathbf{C}$, (3.20), that this sum always reduces to zero. Let $\mathbf{B}_g' = \left[B_{g0}, B_{g1}, ..., B_{gK} \right]$ represent the vector that is the gth row of the matrix $\mathbf{B}$, (3.19). Then an element of $\mathbf{V}$ from the gth row and the g'th column can be expressed as

$$V_{gg'} = A_{gg'} - \mathbf{B}_{g'}' \mathbf{C}^{-1} \mathbf{B}_g \tag{5.32}$$

By adding $\mathbf{B}_g' \mathbf{C}^{-1} \mathbf{C} \mathbf{C}^{-1} \mathbf{B}_{g'}$ and subtracting $\mathbf{B}_g' \mathbf{C}^{-1} \mathbf{B}_{g'}$ to the right-hand side, and algebraically manipulating, (5.32) reduces to

$$V_{gg'} = \sum_{i=1}^{n} \left(\frac{\partial \hat{\theta}}{\partial \eta_0} \right)_i^2 \left(\frac{\partial \zeta_0}{\partial \hat{\theta}} \right)_i \left\{ \left(\frac{\partial \eta_0}{\partial \zeta_0} \right)_i I_i^{(g)} - \mathbf{x}_i' \mathbf{C}^{-1} \mathbf{B}_g \right\} \left\{ \left(\frac{\partial \eta_0}{\partial \zeta_0} \right)_i I_i^{(g')} - \mathbf{x}_i' \mathbf{C}^{-1} \mathbf{B}_{g'} \right\} \tag{5.33}$$

Let $\mathbf{B}_{\Sigma g'} = \sum_{g'=1}^{G} \mathbf{B}_{g'} = \left[\sum_{g'=1}^{G} B_{g'0}, \sum_{g'=1}^{G} B_{g'1}, ..., \sum_{g'=1}^{G} B_{g'K} \right]'$, which is a $K+1$ column vector. Then the sum across the gth row of $\mathbf{V}$ can be written as

$$\sum_{g'=1}^{G} V_{gg'} = \sum_{g=1}^{G} \sum_{i=1}^{n} \left(\frac{\partial \hat{\theta}}{\partial \eta_0} \right)_i^2 \left(\frac{\partial \zeta_0}{\partial \hat{\theta}} \right)_i \left\{ \left(\frac{\partial \eta_0}{\partial \zeta_0} \right)_i I_i^{(g)} - \mathbf{x}_i' \mathbf{C}^{-1} \mathbf{B}_g \right\} \left\{ \left(\frac{\partial \eta_0}{\partial \zeta_0} \right)_i I_i^{(g')} - \mathbf{x}_i' \mathbf{C}^{-1} \mathbf{B}_{g'} \right\}$$

$$= \sum_{i=1}^{n} \left(\frac{\partial \hat{\theta}}{\partial \eta_0} \right)_i^2 \left(\frac{\partial \zeta_0}{\partial \hat{\theta}} \right)_i \left\{ \left(\frac{\partial \eta_0}{\partial \zeta_0} \right)_i I_i^{(g)} - \mathbf{x}_i' \mathbf{C}^{-1} \mathbf{B}_g \right\} \left\{ \left(\frac{\partial \eta_0}{\partial \zeta_0} \right)_i - \mathbf{x}_i' \mathbf{C}^{-1} \mathbf{B}_{\Sigma g'} \right\} \tag{5.34}$$

Substituting in $\mathbf{B}_g = \sum_{i=1}^{n} \left(\partial \hat{\theta} / \partial \eta_0 \right)_i \left(\partial \zeta_0 / \partial \eta_0 \right)_i \mathbf{x}_i$ and $\mathbf{C} = \sum_{i=1}^{n} \left(\partial \hat{\theta} / \partial \eta_0 \right)_i \left(\partial \zeta_0 / \partial \eta_0 \right)_i \mathbf{x}_i \mathbf{x}_i'$, which follow from (5.27), (5.30), and $\mathbf{B}_g' = \left[B_{g0}, B_{g1}, ..., B_{gK} \right]$, this reduces to

$$\sum_{g'=1}^{G} V_{gg'} = \sum_{i=1}^{n} \hat{\theta}_i \left(1 - \hat{\theta}_i \right) I_i^{(g)} \left\{ 1 - \left(\frac{\partial \zeta_0}{\partial \eta_0} \right)_i \mathbf{x}_i' \mathbf{C}^{-1} \mathbf{B}_{\Sigma g'} \right\} \tag{5.35}$$

Under the canonical link function, $\zeta = \eta$, and so (5.35) becomes

$$\sum_{g'=1}^{G} V_{gg'} = \sum_{i=1}^{n} \hat{\theta}_i \left(1-\hat{\theta}_i\right) I_i^{(g)} \left\{1 - \mathbf{x}_i' \mathbf{C}^{-1} \mathbf{B}_{\Sigma g'}\right\} \tag{5.36}$$

We now present Halteman's theorem and proof using our notation.

Theorem 5.1 The sum across any row or down any column of $\mathbf{V}$, (3.13), is equal to zero. (Due to Halteman (1980))

Proof

The matrix $\mathbf{V}$ of (3.13), is the same as (5.14) used in the calculations of $T_{\mathcal{G}}$ when the canonical link function is chosen. A necessary condition for $\mathbf{V}$ to be singular is for the right-hand side of (5.36) to reduce to zero. Partition $\mathbf{C}$ as

$$\mathbf{C} = \begin{bmatrix} \mathbf{G} & \mathbf{H}' \\ \mathbf{H} & \mathbf{J} \end{bmatrix} \tag{5.37}$$

where $\mathbf{G} = C_{00}$, $\mathbf{H} = \left[C_{01}, \quad C_{02}, \quad \ldots \quad C_{0K}\right]'$, and

$$\mathbf{J} = \begin{bmatrix} C_{11} & \ldots & C_{1K} \\ \ldots & \ldots & \ldots. \\ C_{K1} & \ldots & C_{KK} \end{bmatrix} \tag{5.38}$$

The elements of $\mathbf{C}$ where $k, k' = 0,1,\ldots,K$, are then $C_{00} = \sum_{i=1}^{n} \left(\partial\hat{\theta}/\partial\eta_0\right)_i^2 \left(\partial\zeta_0/\partial\hat{\theta}\right)_i x_{i0} x_{i0}$,

$C_{0k} = \sum_{i=1}^{n} \left(\partial\hat{\theta}/\partial\eta_0\right)_i^2 \left(\partial\zeta_0/\partial\hat{\theta}\right)_i x_{i0} x_{ik}$ and $C_{kk'} = \sum_{i=1}^{n} \left(\partial\hat{\theta}/\partial\eta_0\right)_i^2 \left(\partial\zeta_0/\partial\hat{\theta}\right)_i x_{ik} x_{ik'}$.

By Rao (2002),

$$\mathbf{C^{-1}} = \begin{bmatrix} \mathbf{G^{-1}} + \mathbf{M'L^{-1}M} & \mathbf{-M'L^{-1}} \\ \mathbf{-L^{-1}M} & \mathbf{L^{-1}} \end{bmatrix} \tag{5.39}$$

where $\mathbf{L = J - HG^{-1}H'}$ and $\mathbf{M = G^{-1}H}$. Since $\partial\eta_0/\partial\zeta_0 = 1$, the elements of $\mathbf{B}_{\Sigma g'}$ can be expressed as $\mathbf{B}_{\Sigma g'} = \left[\mathbf{G}, \mathbf{H}'\right]'$. Then

$$\mathbf{x}_i' \mathbf{C}^{-1} \mathbf{B}_{\Sigma k'} = \left[1, x_{i1}, \ldots, x_{iK}\right] \begin{bmatrix} \mathbf{G^{-1}} + \mathbf{M'L^{-1}M} & \mathbf{-M'L^{-1}} \\ \mathbf{-L^{-1}M} & \mathbf{L^{-1}} \end{bmatrix} \begin{bmatrix} \mathbf{G} \\ \mathbf{H} \end{bmatrix}$$

$$= \left[1, \mathbf{x}'_{*i}\right] \begin{bmatrix} \mathbf{G}^{-1}\mathbf{G} + \mathbf{M}'\mathbf{L}^{-1}\mathbf{M}\mathbf{G} - \mathbf{M}'\mathbf{L}^{-1}\mathbf{H} \\ -\mathbf{L}^{-1}\mathbf{M}\mathbf{G} + \mathbf{L}^{-1}\mathbf{H} \end{bmatrix}$$

$$= \left[1, \mathbf{x}'_{*i}\right] \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix}$$

$$= 1 \qquad (5.39)$$

where $\mathbf{x}_{*i} = \left[x_{i1}, ..., x_{iK}\right]'$. This reduction occurs since $\mathbf{G}$ is a scalar and so $\mathbf{H} = \mathbf{G}\mathbf{M} = \mathbf{M}\mathbf{G}$.

Thus every summand of (5.36) reduces to zero, and the expression reduces to zero.

Now consider the expression (5.36) under the generalized Tsiatis model

$$\sum_{g'=1}^{G} V_{gg'} = \sum_{i=1}^{n} \hat{\theta}_i \left(1 - \hat{\theta}_i\right) I_i^{(g)} \left\{1 - \left(\frac{\partial \zeta_0}{\partial \eta_0}\right)_i \mathbf{x}_i' \mathbf{C}^{-1} \mathbf{B}_{\Sigma g'}\right\} \qquad (5.39)$$

Under the canonical model, when $\zeta_0 = \eta_0$, (5.39) reduces to (5.36). However, this does not occur when the link is non-canonical, and thus it is possible for (5.39) to sum to something other than zero. Therefore Halteman's proof that $\mathbf{V}$ must be singular under the canonical link does not hold for models with non-canonical links.

5.5 Grouping Method

In the original Tsiatis model, groups are formed by partitioning the covariate space. An alternative is to use a grouping method that is based on the predicted probabilities, such as the deciles-of-risk method. In this method the observed data are first collected, then the predicted probabilities are calculated, ordered, and divided into approximately even groups. These groups are then used in place of a partition of the covariate space. This allows for a more direct comparison of the performance of $T_{\mathcal{G}}$ to HL.

5.6 Examples of Alternative Tsiatis and Generalized Tsiatis Models

The alternative hypothesis of both the original Tsiatis GOF test and the generalized Tsiatis GOF does not provide a specific, fitted alternative model, only the null model under consideration. In addition, because score tests are performed under the null hypothesis, the only model fitted when performing the test is the null model with the linear predictor (5.8). However, to illustrate how the original Tsiatis model and the generalized Tsiatis model might behave if they were fitted to data, we present an example. We generated logistic data ($n = 10000$) using setting 1 from Table 4.1. We then fit a log-log GLM, (5.7), to the data, as well as a generalized Tsiatis model, with log-log link and linear predictor (5.11), and the original Tsiatis model, (3.9), also with log-log link and linear predictor (5.11). The original Tsiatis model had the constant term omitted, since for each observation the indicator functions and the term $x_{i0} = 1$, defined in section 2.1, are linearly dependent. A graph of the fitted probabilities from the three models, as well as those of the true underlying logistic model, each as functions of x_1 are presented in Figure 5.1. The original Tsiatis (with the constant term omitted) and the generalized Tsiatis have similar graphs, with their green and purple lines intersecting over much of the range of x. Both fit the true logistic model more closely than the incorrectly specified null log-log model.

original Tsiatis (with the constant term omitted) and the generalized Tsiatis have similar graphs, with their green and purple lines intersecting over much of the range of x. Both fit the true logistic model more closely than the incorrectly specified null log-log model.

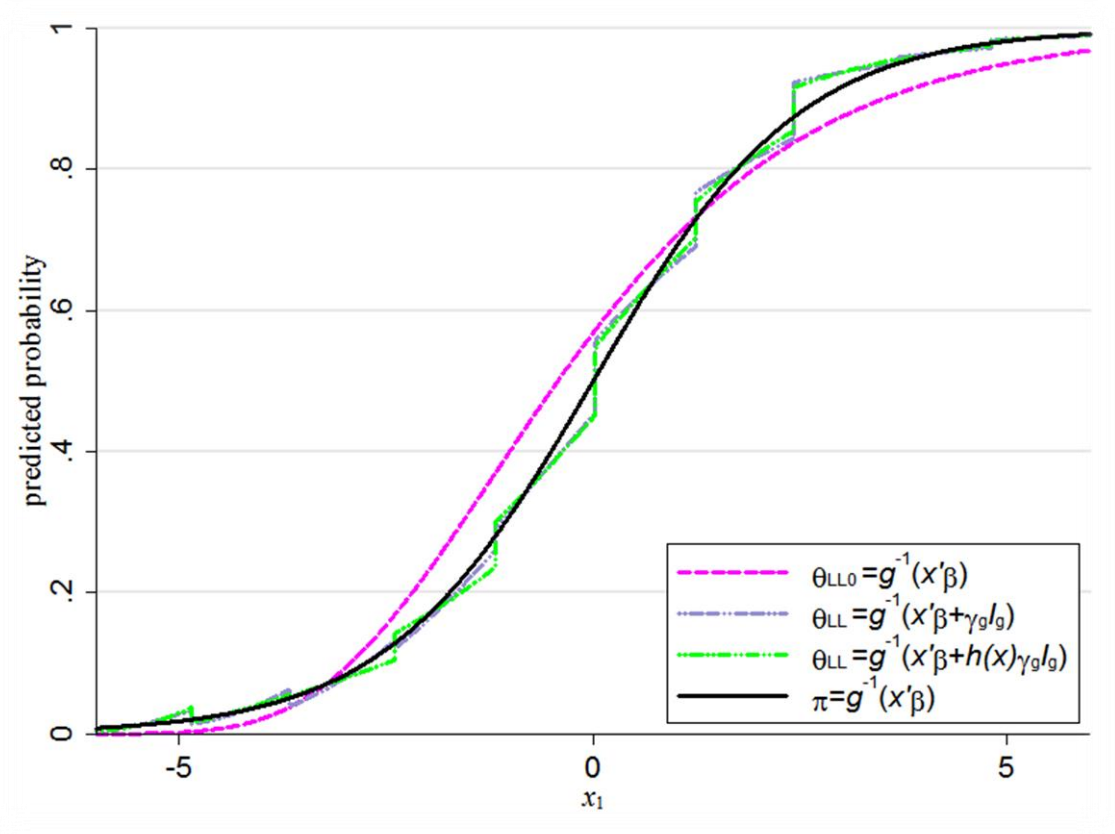

Figure 5.1 Graph of the predicted probabilities of a null log-log model (pink), an original Tsiatis log-log model (purple), a generalized Tsiatis log-log model (green), and the true logistic model (black) as a function of *x*. The data were generated from a logistic model with η=0.8x, *x*~U(-6,6) and *n*=1000. Note that the graphs for the original and generalized Tsiatis models are nearly the same, and both follow the true logistic model's line more closely than the null model does.

5.7 *HL* and J^2 for Binary GLM with Non-Canonical Links

The usual forms of the HL and J^2 statistics are given as (3.6) and (3.23) respectively. As discussed previously in Chapter 4, J^2 can also be expressed as $\mathbf{S}'\mathbf{A}^{-1}\mathbf{S}$. Thus, applying the probit, log-log, complementary log-log, and log link functions to the terms (5.31) and the appropriate term for the elements of $\mathbf{A}$ in Table 5.1, J^2 may be constructed to assess the fit of the corresponding non-canonical model. Similarly, HL may be expressed as (4.1) using a term similar to (4.1), but replacing $\bar{\pi}_g$ with $\bar{\theta}_{\hat{G}g} = \sum_{i\in\psi_g} \hat{\theta}_{\hat{G}i} / n_g$.

5.8 Simulation Study Comparing *HL*, J^2, and T_G Under Non-Canonical Links

5.8.1 Simulation Methods

Simulations were performed to investigate the distributional properties and compare the performances of HL, J^2, and T_G when assessing probit, log-log, complementary log-log and log models under the deciles-of-risk grouping method. Simulations with a large number of replications were performed, to assess whether the distributions of HL and J^2 were approximately $\chi^2(G-2)$, and whether the distribution of T_G was approximately $\chi^2(G)$ under non-canonical link functions. Simulations with fewer replications were performed to examine how well the three statistics under each of the four links maintained test size when a correctly specified model was fitted to generated data, and to compare the power of the statistics to detect a lack of fit when an incorrectly specified model was fit to the data.

Data were generated that varied in the following ways: 1) number of covariates in the model; 2) distributional characteristics of the covariates; 3) inclusion of a quadratic or interaction term; and 4) the number of observations in a sample set. When evaluating the power of each statistic, the ways in which the fitted model departed from the true underlying model were 1) an incorrectly specified link function; 2) omission of a covariate; 3) omission of a quadratic term; or 4) omission of an interaction term. In addition, several settings were based on those used in a study by Blizzard, et al. (2006), who evaluated the performance and distributional characteristics of HL when used to assess log binomial models. Additional settings with other complex models were also studied. A general description of the methods used to perform the simulations is given below, followed by more specific details of the settings studied.

5.8.1.1 General Simulation Methods

Under the settings studied, a model was fit to generated data, predicted probabilities were estimated, and HL, J^2, and T_G (for the specified link) were calculated using the deciles-of-

risk grouping method $(G=10)$. Samples with 500 observations were generated under all of the settings. In addition, samples with only 100 observations were generated under a subset of these settings.

To generate the $(\mathbf{x}, y)$ data, a true underlying model was first chosen with a specific linear predictor, coefficients, distribution of covariates, and a specified link function. Using the chosen setting, a random sample of n covariate vectors $\mathbf{x}$ was then generated for each of r replications of the simulation. Then, to generate outcomes y, probabilities were calculated using (5.7). An exception was when simulations were performed to compare the power of the statistics to detect an incorrectly fitted non-canonical model to data that followed a logistic curve. In this case a logistic probability, $\pi(\mathbf{x}) = \exp(\boldsymbol{\beta}\mathbf{x}')/1+\exp(\boldsymbol{\beta}\mathbf{x}')$, was used to generate the outcomes. The true probabilities were then compared to a value $u \sim U(0,1)$, and the outcome y generated according to the rule $y = I(u < \pi(\mathbf{x}))$, where I is an indicator function such that $I=1$ when the argument is true, and $I=0$ otherwise. All computer simulations described in this chapter were performed using Stata 12 (StataCorp 2012).

5.8.1.2 Investigation of Null Distribution of *HL*, J^2, and $T_{\mathcal{G}}$

Simulations with a large number of replications $(r=100,000)$ were performed to investigate whether HL and J^2 had distributions near $\chi^2(G-2)$, and $T_{\mathcal{G}}$ had a distribution that was approximately $\chi^2(G)$ when evaluating the fit of the non-canonical models studied. All statistics were calculated using the deciles-of-risk grouping method. Data were generated, and then a correctly specified model was fit to the data. Descriptions of the settings used are given in Table 5.2 (a-e). The models studied varied in complexity, and included a model with a single continuous covariate (setting 3), a model with a continuous and a dichotomous covariate (setting 8), and a model with two continuous covariates and an interaction term (setting 21).

5.8.1.3 Empirical Rejection Percentage Under the Null Hypothesis

Simulations with a lower replication number $(r = 10{,}000)$ but a wider range of scenarios were conducted to evaluate how well HL, J^2, and $T_{\mathcal{G}}$ maintained a Type I error rate at the $\alpha = 0.05$ level. By increasing the variety and complexity of the scenarios studied, it was hoped that a deeper understanding of the empirical rejection rates of the statistics could be gained. The settings used to study the null distributions were applied again at this lower replication rate for comparison. Data were generated from models that have linear predictors with the distributional characteristics listed in Table 5.2 (a-e). Datasets of 500 observations were generated for all of these settings, with additional datasets of 100 observations also generated for settings 1, 6, 9, 14, and 20. Correctly specified models were then fitted to these datasets, and the predicted probabilities and test statistics calculated. Finally, HL, J^2, and $T_{\mathcal{G}}$ were compared to the critical value for their respective postulated distributions at the $\alpha = 0.05$ level.

Some of the settings chosen were based on those used in previous studies, thus allowing for the comparison to the results of those studies, which applied HL to logistic and log binomial settings. Settings 1-3 and 8-10 were used by Blizzard, et al. (2006) to evaluate HL as a goodness-of-fit statistic for log binomial models, although they used fewer replications (1000) in their simulations. Settings 1-3 correspond to the first three settings in Blizzard, et al. (2006). These are univariate models with a covariate from a uniform distribution. The slope coefficients were chosen so that $\Pr(Y = 1 \mid x = -6) = 0.01$ and $\Pr(Y = 1 \mid x = -6) = 0.1$ for settings 1 and 2, while $\Pr(Y = 1 \mid x = 0) = 0.3$ for setting 3. Then, given the choice of coefficients that produce these settings, the upper bound of the uniform distributions of settings 1-3 were chosen to be 6, 4 and 2 respectively, so that the marginal probability of response was equal to near 0.2 for settings 1 and 3, and near 0.1 for setting 2. This yielded maximum probabilities for settings 1-3 near 0.99, 0.46 and 0.93 respectively. Thus setting 2 would produce a curve most similar to a logistic, while settings 1 and 3 would be more likely

to have convergence and admissibility problems. For settings 8-10, like Blizzard, et al. (2006), we chose the parameters and coefficients such that under the log link the maximum probabilities were near 0.9, the uniform covariate was a significant predictor, and the log binomial and logistic settings would differ as much as possible. Settings 8 and 9 differ only in the Bernoulli parameter selected, with setting 8 having unbalanced groups and setting 9 balanced. Settings 9, 10 and 11 have balanced groups but the separation between the groups becomes progressively more pronounced.

Settings 1, 4, 5, 13-16, and 17-20 were similar to those used by Hosmer, et al. (1997) who evaluated HL and other goodness-of-fit statistics for logistic models. These were also used in the simulation study presented in 4.5 to study HL, J^2, and T in the logistic setting. The values of the regression coefficients for the non-canonical models differ from those of the logistic setting, but were chosen so that the distributions of the predicted probabilities were comparable to the earlier study. Settings which might produce groups containing only probabilities near 0 or 1 were avoided. Settings 13 to 16 have a linear predictor with a quadratic term, and the regression coefficients are chosen such that the model curve passed through the points $\pi(-1.5) = 0.05$, $\pi(3) = 0.95$, and $\pi(-3) = W$, where $W = 0.01,\ 0.05,\ 0.1$, and 0.4. As the value of W increases, the departure from linearity increases. Settings 17 and 18 were similar to settings in Hosmer, et al. (1997), but the range of the uniform covariates was more narrow. In settings 19 and 20, the linear predictor contains a continuous covariate, a dichotomous covariate and an interaction term. Regression coefficients were chosen such that the model curve passes through the points $\pi(-3,0) = 0.1$, $\pi(-3,1) = 0.1$, $\pi(3,0) = 0.2$, and $\pi(3,1) = 0.3$ or 0.9, with more interaction occurring with the higher maximum probability. Setting 21 also contains an interaction term, but has two continuous covariates rather than a dichotomous and a continuous covariate. Like settings 13-16, settings 22 – 24 also demonstrate increasing departure from linearity, but in these cases, the settings contain a dichotomous covariate along with a continuous covariate and quadratic term. Additional settings studied included other models with two or three covariates (settings 6, 7). All settings used in the power simulations were included in the null simulations.

All of the settings studied were applied to the probit, log-log, complementary log-log and log binomial models. The range of the probabilities for all of these models is limited to between 0 and 1, with the exception of the log binomial model. Because the log link can result in probabilities greater than 1, there can be difficulties fitting the model parameters and results may be inadmissible. Therefore, in addition to the specifications for the settings described above, the settings were also chosen so that these problems would not occur frequently in the log binomial simulations. Table 5.2(a-e) contains the specifications of the settings for the simulations used to investigate the null distributions of *HL*, J^2 , and $T_{\mathcal{G}}$, as well as the power of each statistic to detect an incorrectly specified link function.

Table 5.2 (a-e) Settings for simulations used to investigate the null distributions of *HL*, J^2 , and $T_{\mathcal{G}}$, and to evaluate the power of each statistic to detect an incorrectly specified link function.

Table 5.2a. Linear predictor characteristics for settings 1-24 for all links.

		Distribution of covariate		
Setting	Linear predictor	x_1	x_2	x_3
1	$\beta_0+\beta_1 x_1$	U(-6,6)	.	.
2	$\beta_0+\beta_1 x_1$	U(-6,4)	.	.
3	$\beta_0+\beta_1 x_1$	U(-6,2)	.	
4	$\beta_0+\beta_1 x_1$	$\chi^2(4)$	.	.
5	$\beta_0+\beta_1 x_1$	N(0,1.5)	.	.
6	$\beta_0+\beta_1 x_1+\beta_2 x_2$	U(-6,0)	N(3,1)	.
7	$\beta_0+\beta_1 x_1+\beta_2 x_2$	U(-6,0)	$\chi^2(4)$	.
8	$\beta_0+\beta_1 x_1+\beta_2 x_2$	Ber(0.2)	U($2x_1$-6,$2x_1$+2)	.
9	$\beta_0+\beta_1 x_1+\beta_2 x_2$	Ber(0.5)	U($2x_1$-6,$2x_1$+2)	.
10	$\beta_0+\beta_1 x_1+\beta_2 x_2$	Ber(0.5)	U($2x_1$-6,$2x_1$+2)	.
11	$\beta_0+\beta_1 x_1+\beta_2 x_2$	Ber(0.5)	U($2x_1$-6,$2x_1$+2)	.
12	$\beta_0+\beta_1 x_1+\beta_2 x_2$	Ber(0.5)	U($2x_1$-6,$2x_1$+2)	.
13	$\beta_0+\beta_1 x_1+\beta_2 x_1^2$	U(-3,3)	x_1^2	.
14	$\beta_0+\beta_1 x_1+\beta_2 x_1^2$	U(-3,3)	x_1^2	.
15	$\beta_0+\beta_1 x_1+\beta_2 x_1^2$	U(-3,3)	x_1^2	.
16	$\beta_0+\beta_1 x_1+\beta_2 x_1^2$	U(-3,3)	x_1^2	.
17	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3$	U(-1,1)	U(-1,1)	U(-1,1)
18	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3$	U(-1,1)	N(0,1.5)	$\chi^2(4)$
19	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2$	U(-3,3)	Ber(0.5)	$x_1 x_2$
20	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2$	U(-3,3)	Ber(0.5)	$x_1 x_2$
21	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2$	U(-1,1)	N(0,1.5)	$x_1 x_2$
22	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1^2$	U(-3,3)	Ber(0.5)	x_1^2
23	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1^2$	U(-3,3)	Ber(0.5)	x_1^2
24	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1^2$	U(-3,3)	Ber(0.5)	x_1^2

Table 5.2b. The regression coefficients for the probit models and the distributional characteristics of the probit probabilities.

	Regression coefficients				Distribution Characteristics of θ_{Pr}†				
Setting	β_0	β_1	β_2	β_3	θ*(1)*	*Q1*	*Q2*	*Q3*	θ*(N)*
1	-0.20	0.38	.	.	0.01	0.09	0.44	0.86	0.98
2	-0.20	0.38	.	.	0.01	0.07	0.30	0.69	0.91
3	0.40	0.38	.	.	0.03	0.13	0.38	0.71	0.90
4	-2.20	0.18	.	.	0.02	0.04	0.06	0.12	0.98
5	-0.50	0.32	.	.	0.01	0.20	0.31	0.45	0.87
6	0.00	0.40	0.40	.	0.01	0.28	0.54	0.79	0.97
7	-0.50	0.20	0.05	.	0.03	0.08	0.16	0.25	0.75
8	-0.10	0.41	0.18	.	0.15	0.25	0.38	0.50	0.79
9	-0.10	0.41	0.18	.	0.13	0.32	0.45	0.60	0.83
10	-0.10	0.69	0.10	.	0.25	0.38	0.56	0.70	0.81
11	-0.10	0.99	0.01	.	0.43	0.49	0.76	0.80	0.83
12	-0.10	0.18	0.10	.	0.27	0.37	0.44	0.50	0.63
13	-0.83	0.60	0.03	.	0.02	0.06	0.23	0.63	0.92
14	-1.23	0.49	0.12	.	0.04	0.05	0.14	0.48	0.88
15	-1.45	0.43	0.16	.	0.04	0.05	0.08	0.38	0.88
16	-2.04	0.26	0.28	.	0.01	0.02	0.09	0.30	0.88
17	0.00	0.27	0.27	0.27	0.27	0.26	0.44	0.52	0.61
18	-1.90	0.12	0.12	0.10	0.02	0.05	0.07	0.10	0.68
19	-1.06	0.07	0.16	0.06	0.08	0.14	0.16	0.17	0.31
20	-1.06	0.07	1.10	0.36	0.07	0.15	0.16	0.42	0.87
21	-0.50	0.10	0.10	0.10	0.11	0.26	0.27	0.30	0.51
22	-1.60	0.49	0.50	0.11	0.02	0.05	0.09	0.37	0.91
23	-1.90	0.38	0.50	0.18	0.01	0.03	0.08	0.26	0.91
24	-2.20	0.28	0.50	0.26	0.01	0.04	0.08	0.26	0.92

† Expected values of the smallest, largest, and three quartiles of the resulting distribution of the probit probabilities for a sample size of 500.

Table 5.2c. The regression coefficients for the log-log models, and the distributional characteristics of the log-log probabilities.

	Regression coefficients				Distribution Characteristics of θ_{LL}†				
Setting	β_0	β_1	β_2	β_3	θ*(1)*	*Q1*	*Q2*	*Q3*	θ*(N)*
1	0.80	0.38	.	.	0.02	0.27	0.66	0.89	0.96
2	0.80	0.38	.	.	0.02	0.20	0.52	0.78	0.89
3	1.70	0.57	.	.	0.01	0.19	0.57	0.84	0.94
4	-1.20	0.18	.	.	0.03	0.07	0.13	0.25	0.96
5	0.50	0.32	.	.	0.01	0.42	0.54	0.66	0.89
6	0.80	0.40	0.40	.	0.02	0.45	0.66	0.81	0.93
7	0.40	0.20	0.05	.	0.12	0.25	0.38	0.48	0.76
8	0.20	0.41	0.18	.	0.14	0.24	0.37	0.47	0.70
9	0.20	0.41	0.18	.	0.11	0.30	0.42	0.55	0.73
10	0.20	0.69	0.10	.	0.27	0.38	0.53	0.64	0.72
11	0.20	0.99	0.01	.	0.41	0.47	0.69	0.72	0.76
12	0.20	0.18	0.10	.	0.26	0.36	0.43	0.50	0.60
13	-0.32	0.63	0.08	.	0.02	0.05	0.25	0.69	0.92
14	-0.53	0.56	0.12	.	0.04	0.05	0.21	0.64	0.9
15	-0.83	0.51	0.17	.	0.04	0.06	0.11	0.55	0.9
16	-1.31	0.36	0.28	.	0.02	0.03	0.15	0.45	0.91
17	0.50	0.27	0.27	0.27	0.30	0.51	0.59	0.66	0.81
18	-0.70	0.12	0.12	0.10	0.02	0.18	0.26	0.31	0.67
19	-0.66	0.06	0.15	0.05	0.07	0.14	0.17	0.2	0.3
20	-0.66	0.06	1.36	0.45	0.08	0.16	0.19	0.56	0.87
21	-0.20	0.10	0.10	0.10	0.11	0.25	0.27	0.31	0.56
22	-0.90	0.49	0.50	0.11	0.02	0.06	0.16	0.54	0.86
23	-1.20	0.38	0.50	0.18	0.02	0.07	0.12	0.42	0.84
24	-1.30	0.28	0.50	0.26	0.02	0.08	0.19	0.44	0.89

† Expected values of the smallest, largest, and three quartiles of the resulting distribution of the log-log probabilities for a sample size of 500.

Table 5.2d. The regression coefficients for the complementary log-log models, and the distributional characteristics of the complementary log-log probabilities.

	Regression coefficients				Distribution Characteristics of θ_{Cll}†				
Setting	β_0	β_1	β_2	β_3	θ*(1)*	*Q1*	*Q2*	*Q3*	θ*(N)*
1	-1.10	0.38	.	.	0.04	0.11	0.30	0.70	0.96
2	-2.00	0.38	.	.	0.02	0.04	0.10	0.26	0.49
3	-0.30	0.57	.	.	0.03	0.08	0.22	0.56	0.89
4	-4.00	0.18	.	.	0.02	0.03	0.03	0.05	0.78
5	-1.20	0.32	.	.	0.03	0.18	0.26	0.35	0.81
6	-1.20	0.40	0.40	.	0.02	0.15	0.29	0.50	0.89
7	-1.30	0.20	0.05	.	0.05	0.09	0.14	0.20	0.67
8	-0.60	0.41	0.18	.	0.20	0.27	0.37	0.47	0.76
9	-0.60	0.41	0.18	.	0.17	0.31	0.42	0.55	0.80
10	-0.60	0.69	0.10	.	0.28	0.37	0.52	0.64	0.76
11	-0.60	0.99	0.01	.	0.39	0.44	0.71	0.76	0.80
12	-0.60	0.18	0.10	.	0.28	0.36	0.41	0.48	0.60
13	-1.52	0.91	-0.04	.	0.01	0.06	0.22	0.61	0.92
14	-2.44	0.63	0.15	.	0.04	0.05	0.09	0.32	0.87
15	-2.86	0.51	0.24	.	0.04	0.05	0.08	0.21	0.88
16	-3.75	0.25	0.43	.	0.02	0.03	0.07	0.21	0.93
17	-0.30	0.27	0.27	0.27	0.33	0.48	0.55	0.64	0.84
18	-1.40	0.12	0.12	0.10	0.10	0.23	0.29	0.34	0.61
19	-1.88	0.13	0.24	0.08	0.08	0.14	0.16	0.19	0.32
20	-1.88	0.13	1.17	0.39	0.07	0.15	0.17	0.34	0.90
21	-1.50	0.10	0.10	0.10	0.06	0.16	0.18	0.21	0.57
22	-2.20	0.49	0.50	0.11	0.08	0.10	0.15	0.37	0.84
23	-2.60	0.38	0.50	0.18	0.07	0.09	0.12	0.29	0.82
24	-3.20	0.28	0.50	0.26	0.03	0.05	0.09	0.18	0.83

† Expected values of the smallest, largest, and three quartiles of the resulting distribution of the complementary log-log probabilities for a sample size of 500.

Table 5.2e. The regression coefficients for the log binomial models, and the distributional characteristics of the log probabilities.

	Regression coefficients				Distribution Characteristics of $\theta_{LB\dagger}$				
Setting	β_0	β_1	β_2	β_3	θ*(1)*	*Q1*	*Q2*	*Q3*	θ*(N)*
1	-2.30	0.38	.	.	0.01	0.03	0.10	0.32	0.95
2	-2.30	0.38	.	.	0.01	0.03	0.07	0.18	0.46
3	-1.20	0.57	.	.	0.01	0.03	0.10	0.30	0.90
4	-2.60	0.13	.	.	0.08	0.10	0.12	0.15	0.63
5	-1.20	0.20	.	.	0.13	0.25	0.30	0.37	0.73
6	-2.20	0.40	0.40	.	0.02	0.06	0.11	0.21	0.75
7	-1.10	0.20	0.05	.	0.11	0.16	0.22	0.30	0.62
8	-1.20	0.41	0.18	.	0.11	0.16	0.24	0.35	0.88
9	-1.20	0.41	0.18	.	0.11	0.21	0.31	0.46	0.90
10	-1.20	0.69	0.10	.	0.17	0.25	0.38	0.60	0.89
11	-1.20	0.99	0.01	.	0.28	0.29	0.53	0.81	0.85
12	-1.20	0.18	0.10	.	0.17	0.23	0.30	0.37	0.54
13	-1.71	0.75	-0.07	.	0.01	0.05	0.18	0.48	0.88
14	-2.51	0.48	0.11	.	0.04	0.05	0.08	0.22	0.86
15	-2.86	0.37	0.18	.	0.05	0.05	0.07	0.16	0.85
16	-3.55	0.14	0.34	.	0.03	0.04	0.06	0.16	0.83
17	-0.80	0.27	0.27	0.27	0.23	0.37	0.45	0.55	0.88
18	-2.20	0.12	0.12	0.10	0.07	0.13	0.16	0.20	0.65
19	-1.96	0.12	0.20	0.07	0.08	0.12	0.15	0.19	0.31
20	-1.96	0.12	0.75	0.25	0.09	0.13	0.17	0.30	0.88
21	-1.40	0.10	0.10	0.10	0.12	0.22	0.24	0.28	0.55
22	-3.11	0.49	0.50	0.11	0.02	0.04	0.06	0.16	0.80
23	-3.40	0.38	0.50	0.18	0.03	0.04	0.06	0.12	0.80
24	-3.80	0.28	0.50	0.26	0.02	0.03	0.05	0.10	0.80

† Expected values of the smallest, largest, and three quartiles of the resulting distribution of the log binomial probabilities for a sample size of 500.

5.8.1.4 Power

Simulations were performed to assess the power of HL, J^2, and $T_{\mathcal{G}}$ to detect a departure from a true underlying model under each of the non-canonical links. Data were generated using some of the settings of Table 5.2(a-e), and then a model was fitted to the data using an incorrectly specified linear predictor, which had either a covariate, a quadratic term or an interaction term omitted. In the second set of power analyses, log binomial models were fitted to data generated from the logistic settings described in Table 5.3. This included individual settings (25-27, 36-37), as well as three series of related settings. In one series, the dichotomous term has increasing influence (settings 28-30). In the second series, the non-

linearity of the model increases (settings 31-35). In the third series, the interaction between terms in the model increases (settings 38-40). All simulations were performed with a replication number of $r = 10,000$. A dataset of 500 observations were generated for each setting. In addition, datasets with 100 observations were generated for settings 6, 14, and 20.

5.8.2 Simulation Results

5.8.2.1 Distribution of *HL*, J^2, and $T_{\mathcal{G}}$ Under Non-Canonical Link Functions

Simulations were performed with a large number of replications $r = 100,000$ under settings 3, 10, and 21 described in Table 5.2 (a-e) for each of the non-canonical links studied. Histograms of the values observed that were less than 32 are presented in Figure 5.2 through Figure 5.13. Summary statistics for all observations are presented in Table 5.4.

The statistics were each compared to the critical value of their postulated distributions. The asymptotic distribution of HL under the logit link was reported by Hosmer and Lemeshow to be $\chi^2(G-2)$. Pigeon, et al. (1999b) state that J^2 has an asymptotic distribution that is approximately $\chi^2(G-1)$, which is not dependent on grouping method. However, the results of the simulation study described in section 4.5 indicate that, when the deciles-of-risk grouping method is used, the distribution of J^2 under the logit link is close to $\chi^2(G-2)$. Thus, with the number of groups set at 10, the values of HL and J^2 were both compared to the critical value of $\chi^2(8)$ at the $\alpha = 0.05$ level. Based on section 5.4, the values of $T_{\mathcal{G}}$ for all links were compared to the critical value of $\chi^2(10)$ at the $\alpha = 0.05$ level, which differs from the critical value derived from a $\chi^2(G-1)$ used under the original Tsiatis statistic in the logistic setting. The rank of $\mathbf{V}$ was recorded for each replication, and was observed to be 10 in all cases.

The means of HL and J^2 were significantly different from 8, and likewise $T_{\mathcal{G}}$ from 10, at the five per cent level, if they fell outside of the intervals (7.98, 8.03) and (9.97, 10.03)

Table 5.3 The distributional characteristics of the covariates and the linear predictors of the settings used to examine the power of *HL*, J^2, and *T*, to detect an incorrectly specified link function.

		Distribution of covariate			Logistic Coefficients				Fitted Log binomial Pr (n=500) Distribution characteristics				
Setting	Linear predictor	x_1	x_2	x_3	β_0	β_1	β_2	β_3	$\theta(1)$	*Q1*	*Q2*	*Q3*	$\theta(N)$
25	$\beta_0+\beta_1x_1$	U(-6,6)	.	.	-0.80	0.38	.	.	0.04	0.13	0.31	0.58	0.81
26	$\beta_0+\beta_1x_1$	U(-6,6)	.	.	-2.30	0.38	.	.	0.01	0.03	0.09	0.24	0.49
27	$\beta_0+\beta_1x_1+\beta_2x_2$	Ber(0.2)	U($2x_1$-6,$2x_1$+2)	.	-1.20	0.41	0.55	.	0.01	0.04	0.12	0.29	0.80
28	$\beta_0+\beta_1x_1+\beta_2x_2$	Ber(0.5)	U($2x_1$-6,$2x_1$+2)	.	-1.20	0.41	0.55	.	0.01	0.07	0.18	0.39	0.80
29	$\beta_0+\beta_1x_1+\beta_2x_2$	Ber(0.5)	U($2x_1$-6,$2x_1$+2)	.	-1.20	0.69	0.48	.	0.02	0.09	0.21	0.41	0.80
30	$\beta_0+\beta_1x_1+\beta_2x_2$	Ber(0.5)	U($2x_1$-6,$2x_1$+2)	.	-1.20	0.90	0.50	.	0.02	0.09	0.22	0.44	0.84
31	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2	.	-1.40	1.00	-0.02	.	0.01	0.05	0.20	0.51	0.80
32	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2	.	-2.22	0.72	0.16	.	0.05	0.05	0.10	0.32	0.80
33	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2		-2.59	0.60	0.24	.	0.05	0.06	0.08	0.24	0.79
34	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2	.	-3.00	0.46	0.33	.	0.04	0.05	0.08	0.19	0.79
35	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2	.	-3.49	0.30	0.44	.	0.03	0.04	0.07	0.21	0.79
36	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$	U(-1,1)	U(-1,1)	U(-1,1)	-0.40	0.63	0.63	0.63	0.12	0.30	0.40	0.51	0.77
37	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$	U(-1,1)	N(0,1.5)	$\chi^2(4)$	-1.40	0.25	0.25	0.12	0.09	0.22	0.28	0.36	0.72
38	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	U(-3,3)	Ber(0.5)	x_1x_2	-1.79	0.14	0.27	0.09	0.10	0.12	0.16	0.19	0.30
39	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	U(-3,3)	Ber(0.5)	x_1x_2	-1.79	0.14	0.90	0.30	0.10	0.13	0.17	0.29	0.60
40	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	U(-3,3)	Ber(0.5)	x_1x_2	-1.79	0.14	1.56	0.52	0.10	0.13	0.18	0.44	0.85

respectively. The percentage of observations that were larger than the critical values for the corresponding chi-square distributions at the $\alpha = 0.01$, 0.05, and 0.10 levels were significantly different from the percentages expected if they fell outside of the 95% confidence intervals (0.94, 1.06), (4.87, 5.14) and (9.81, 10.19) respectively. The standard deviation, skewness and kurtosis of $\chi^2(8)$ are 4, 1 and 4.5 respectively, while for $\chi^2(10)$ they are 4.47, 0.89 and 4.2 respectively.

The histograms for the values of the three statistics calculated under the univariate model of setting 3 for each of the four non-canonical links studied are shown in Figure 5.2 through Figure 5.5. When the probit, log-log, and complementary log-log links were used, the histograms of all three statistics followed the curve of their theoretical distributions closely. However, when the log link was used, the histograms had peaks that were slightly higher than those of the respective postulated distributions of the statistics, $\chi^2(8)$ and $\chi^2(10)$. The Kolmogorov-Smirnov tests, reported in Table 5.4 evaluate the null hypothesis that the probability distributions of the statistics are the same as their theoretical distributions, against the alternative that the distributions are different. Note that with very large samples sizes, the test may be overpowered and significant differences may be considered unimportant. The largest observed difference is given in parentheses. Our results indicated that the distributions of HL and J^2 were significantly different from $\chi^2(8)$ under all links, and that the distribution of $T_{\mathcal{G}}$ is significantly different than $\chi^2(10)$ under all links. Note that the largest differences observed for HL and J^2 were very similar. For setting 3, under the log link, the mean, skewness and kurtosis values of HL and J^2 were higher than expected, and there were higher percentages of observations above all of the percentiles studied. The skewness and kurtosis of T_{LB} were also higher than expected, but the mean was lower, and there were fewer than expected observations above the 90th percentile. Under the probit link, the means of J^2 and T_{Pr} were higher than expected. Under the log-log link, the mean of J^2 and the kurtosis of all three statistics were high. For both J^2 and T_{LL}, there were fewer than expected observations above the 90th and 95th percentiles. Under the complementary log-log

link, the mean of HL was high, as was the kurtosis of both HL and J^2. For both HL and T_{Cll}, there were fewer than expected observations above the 90th percentile.

The histograms for setting 10, displayed in Figure 5.6 through Figure 5.9 , indicate that the histograms of $T_{\mathcal{G}}$ for all links again followed the $\chi^2(10)$ curve closely, while in contrast, the histograms of HL and J^2 for all links were shifted to the left and had higher peaks than that of the $\chi^2(8)$ curve. The Kolmogorov-Smirnov tests, reported in Table 5.4, indicated that the distributions of both HL and J^2 were significantly different from $\chi^2(8)$ under all links, while the distribution of $T_{\mathcal{G}}$ was significantly different from $\chi^2(10)$ under all links. The largest differences observed were small and relatively unimportant. The mean and number of observations above the specified percentiles for HL and J^2 were much lower than expected for all links. Standard deviation values were also low, while skewness and kurtosis were close to the values expected. For $T_{\mathcal{G}}$, the means under all of the links were slightly higher than expected. All other summary statistics were near the values expected except for the skewness and kurtosis of T_{Cll}, which were high. This was driven by a single large value of T_{Cll} near 143. When this one value was removed, the skewness dropped to 0.86 and the kurtosis to 4.17.

The histograms of the statistics under setting 21 for each of the four non-canonical links studied are displayed in Figure 5.10 through Figure 5.13. The histograms of $T_{\mathcal{G}}$ for all links followed the $\chi^2(10)$ curve closely. The histograms of HL and J^2 followed the $\chi^2(8)$ curve fairly closely, except in the case of the complementary log-log, where the histograms had a somewhat higher peak and was shifted left. The Kolmogorov-Smirnov tests, reported in Table 5.4, indicated that the distribution of both HL and J^2 were significantly different than $\chi^2(8)$ under all links, and that the distribution of $T_{\mathcal{G}}$ was significantly different than $\chi^2(10)$ under all links. The largest differences observed were small and relatively unimportant. The mean, standard deviation, and the percentage of observations above the

percentiles studied were lower than expected for both HL and J^2 under all links, while the values of skewness and kurtosis were near those expected. The means of $T_{\mathcal{G}}$ were higher than expected for all links. There were also higher numbers of observations above the 90th and 95th percentiles for T_{LL}, and higher numbers of observations above the 90th percentile of $T_{\Pr}$.

Under all of the links and all three settings, the three statistics agree on whether or not to reject the null hypothesis approximately 97% of the time. In most cases, when they did disagree, HL and J^2 agreed with each other and disagreed with $T_{\mathcal{G}}$

5.8.2.2 Empirical Rejection Percentage Under the Null Hypothesis

Simulations were performed with a smaller number of replications $(r = 10,000)$ under the settings listed in Table 5.2 for each non-canonical model. The values of HL and J^2 were compared to the critical value of $\chi^2(8)$ at the $\alpha = 0.05$ level, while the values of $T_{\mathcal{G}}$ were compared to the critical value of $\chi^2(10)$ at the $\alpha = 0.05$ level. Summary statistics and the empirical null rejection percentages of HL, J^2 and $T_{\mathcal{G}}$ from the simulations are presented in Table 5.5 (a-d). The means of HL and J^2 were significantly different from eight, and likewise $T_{\mathcal{G}}$ from ten, at the five per cent level if they fell outside of the intervals $(7.92,8.02)$ and $(9.91,10.09)$ respectively. Rejection percentages were significantly different from five per cent at the five per cent level if they fell outside of the interval $(4.57,5.43)$.

Overall, when the number of observations was 500, the null Type I error rate of $T_{\mathcal{G}}$ was consistently well maintained. In contrast, HL and J^2 both had rates that were often lower than expected, but also had settings where rejection percentages were higher than expected. The results of the simulations indicate that the rejection percentages of HL and J^2 were

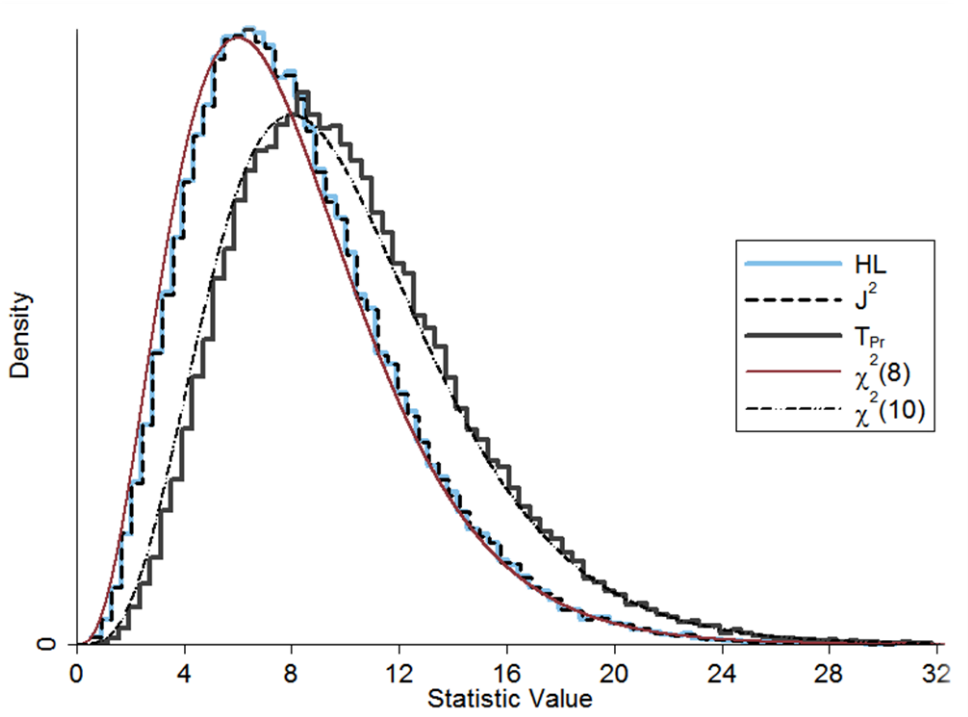

Figure 5.2 Histogram of 100,000 replications of setting 3 using the probit link. (η=0.4 + 0.38x1, x1~U(-6,2), and n=500) The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

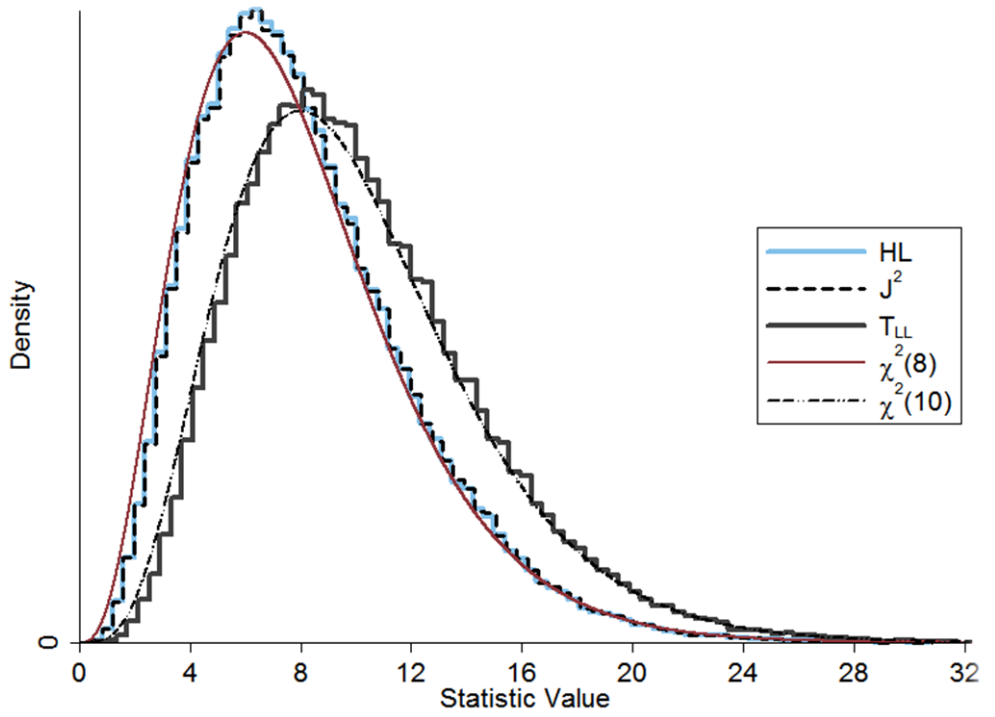

Figure 5.3 Histogram of 100,000 replications of setting 3 using the log-log link. (η=1.7+ 0.57x1, x1~U(-6,2), and n=500) The probability density function curves for $\chi2(8)$ and $\chi2(10)$ are included for comparison.

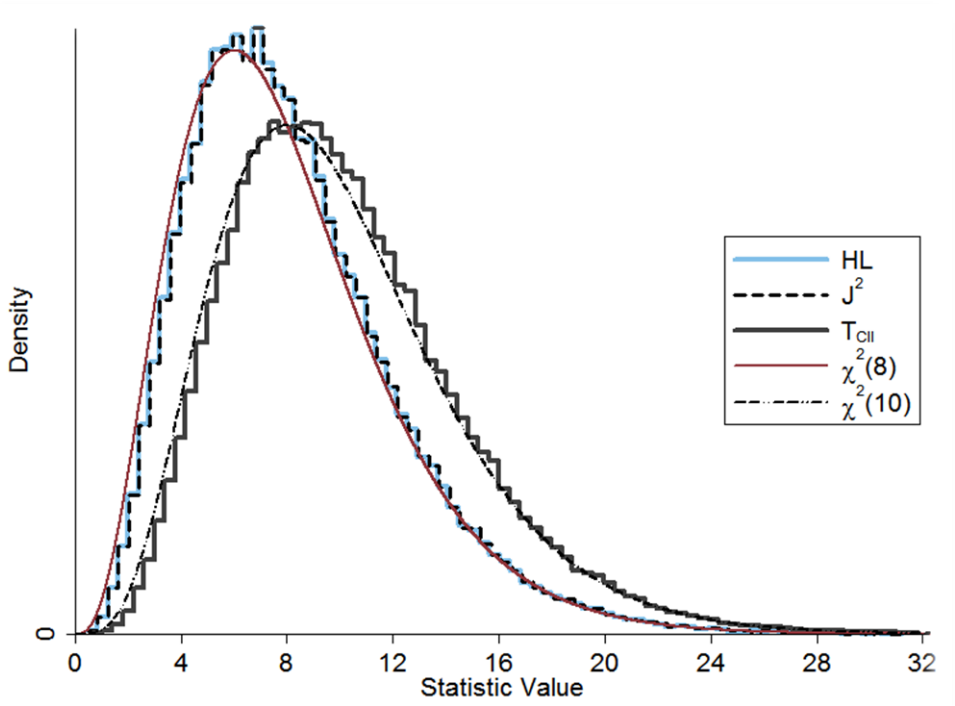

Figure 5.4 Histogram of 100,000 replications of setting 3 using the complementary log-log link. (η=-0.3 + 0.57x_1, x_1~U(-6,2), and n=500).The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

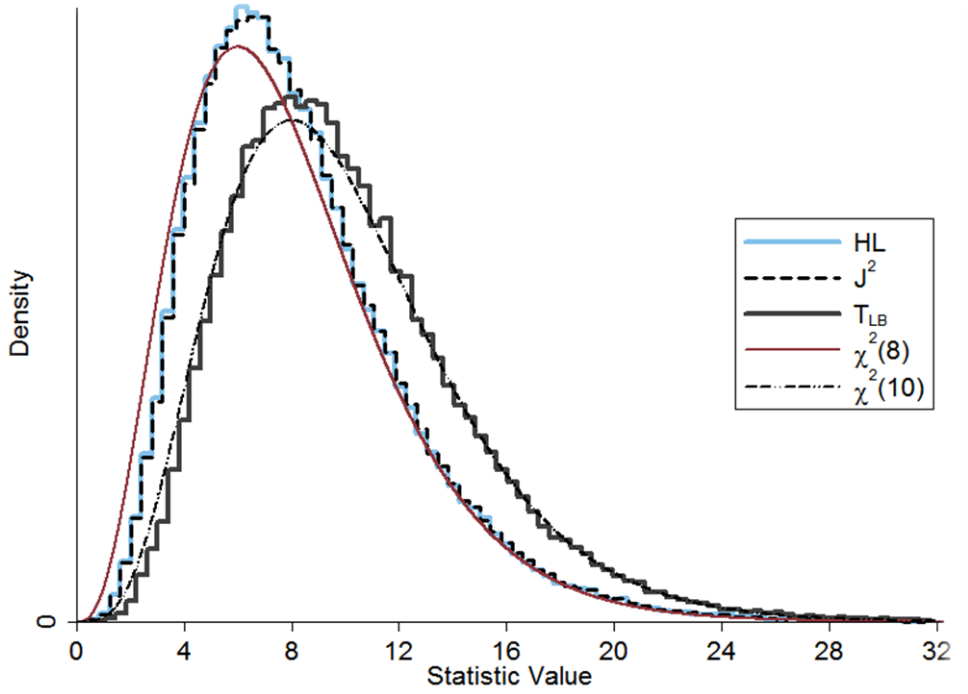

Figure 5.5 Histogram of 100,000 replications of setting 3 using the log link. (η=-1.2 + 0.57x_1, x_1~U(-6,2), and n=500) The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

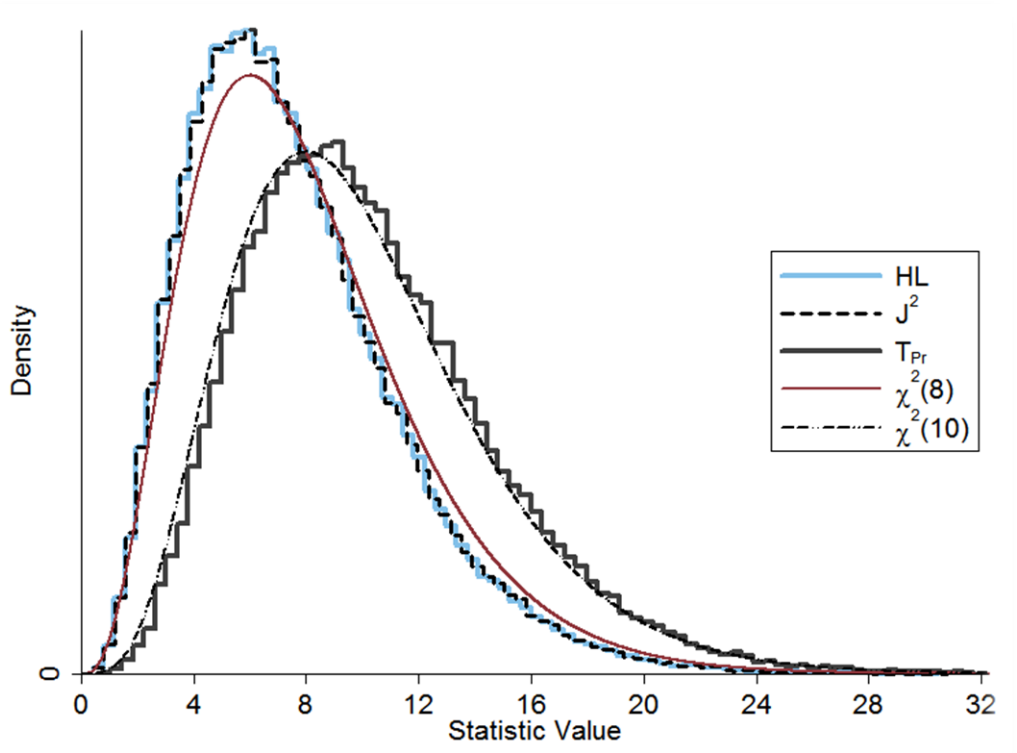

Figure 5.6 Histogram of 100,000 replications of setting 10 using the probit link. (η=-0.1+0.69x_1+0.1x_2, x_1~Bern(0.5), x_2~U(2 x_1-6,2 x_1+2), and n=500) The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

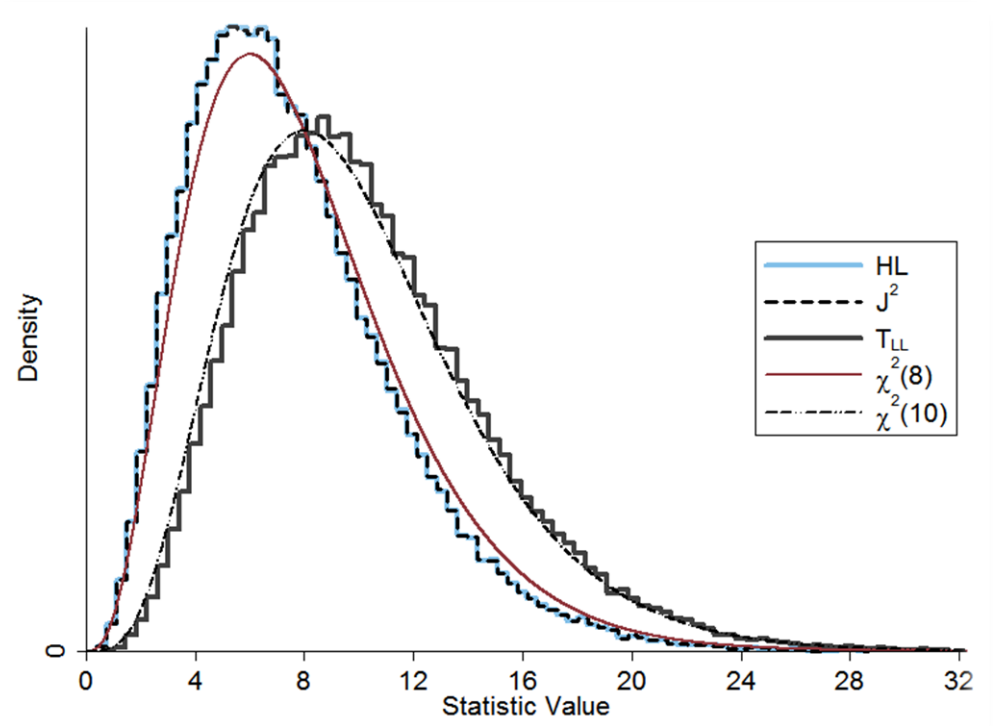

Figure 5.7 Histogram of 100,000 replications of setting 10 using the log-log link. (η=0.2+0.69x_1+0.1x_2, x_1~Bern(0.5), x_2~U(2 x_1-6,2 x_1+2), and n=500) The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

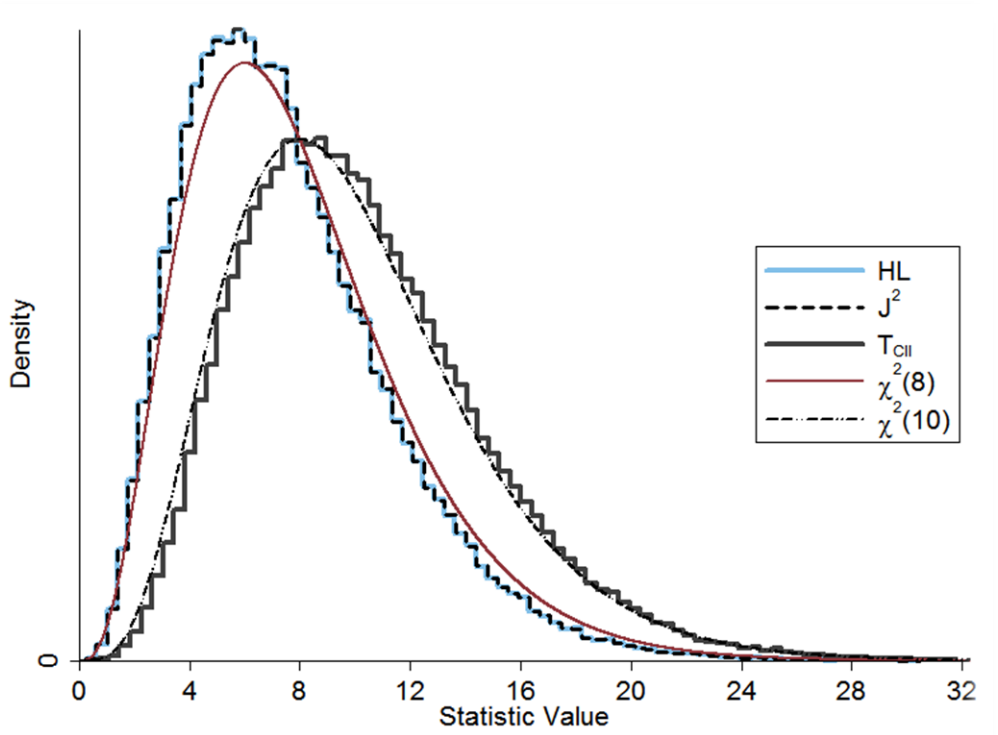

Figure 5.8 Histogram of 100,000 replications of setting 10 using the complementary log- log link. (η=-0.6+0.69x_1+0.1x_2, x_1~Bern(0.5), x_2~U(2 x_1-6,2 x_1+2), and n=500) The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

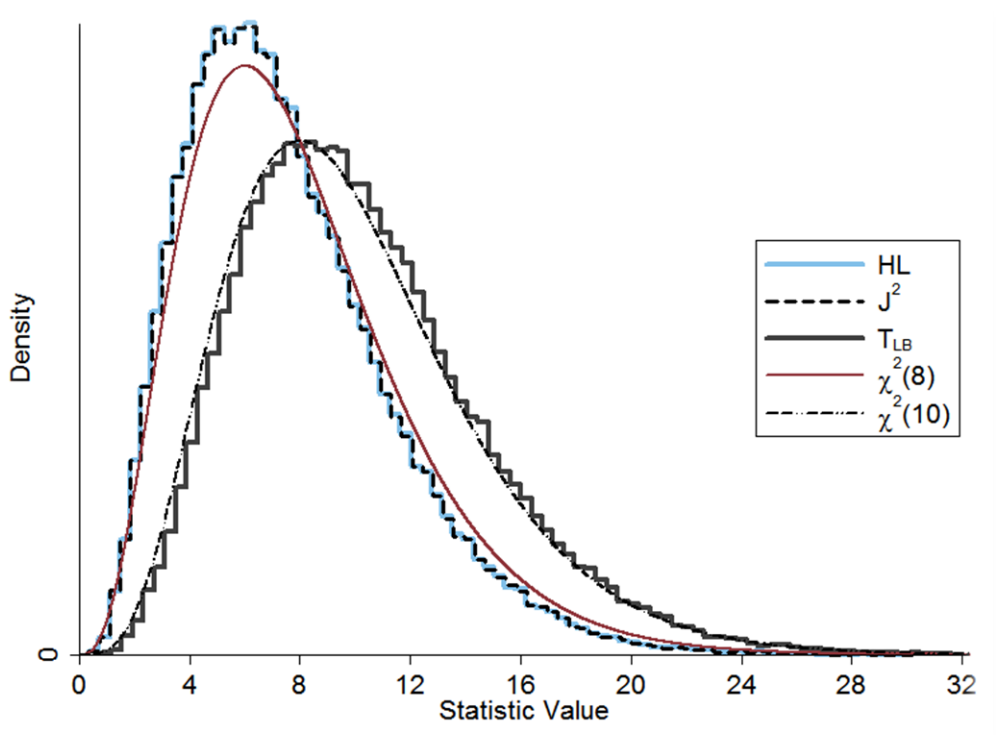

Figure 5.9 Histogram of 100,000 replications of setting 10 using the log link. (η = -1.2+0.69x_1+0.1x_2, x_1~Bern(0.5), x_2~U(2 x_1-6,2 x_1+2), and n=500) The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

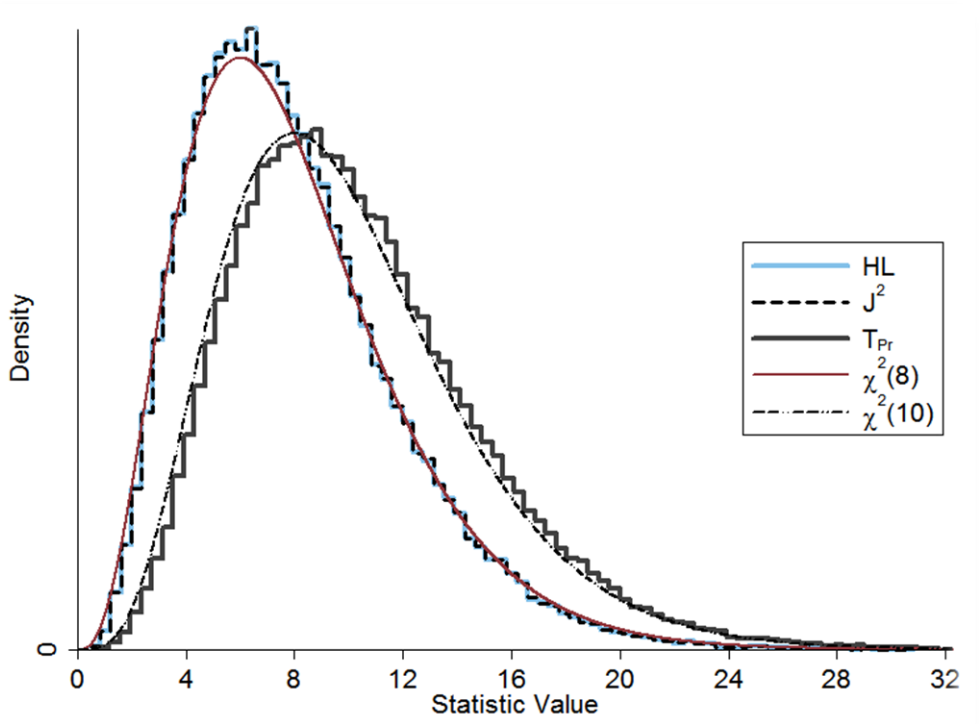

Figure 5.10 Histogram of 100,000 replications of setting 21 using the probit link. (η=-0.5+0.1x_1+0.1x_2, +0.1 x_1x_2, x_1~U(-1,1), x_2~N(0,1.5), and n=500)The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

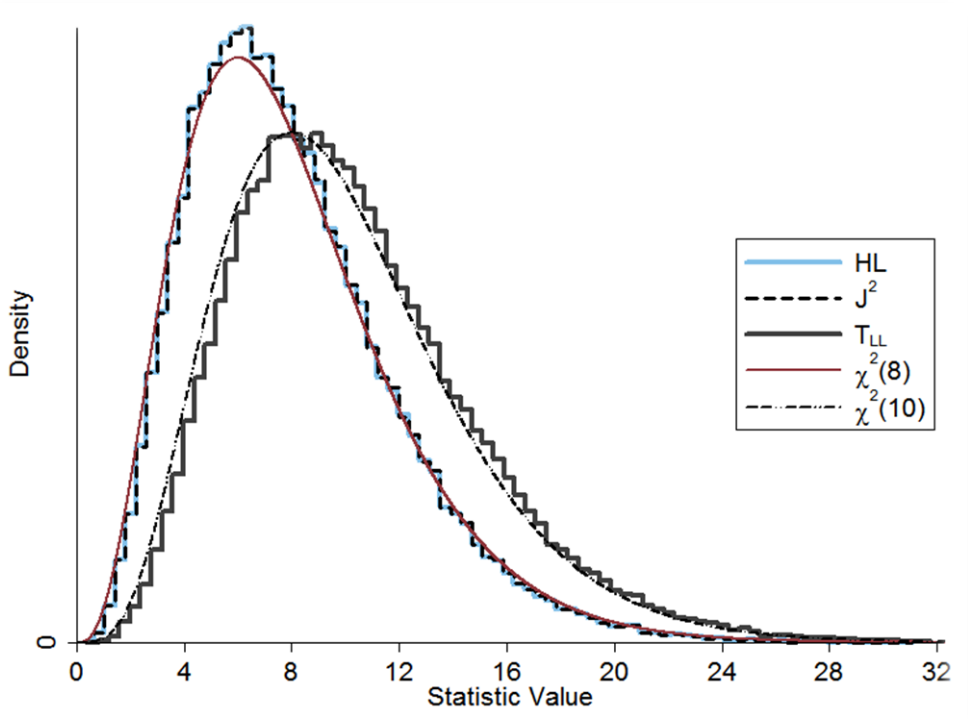

Figure 5.11 Histogram of 100,000 replications of setting 21 using the log-log link. (η=-0.2+0.1x_1+0.1x_2, +0.1 x_1x_2, x_1~U(-1,1), x_2~N(0,1.5), and n=500) The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

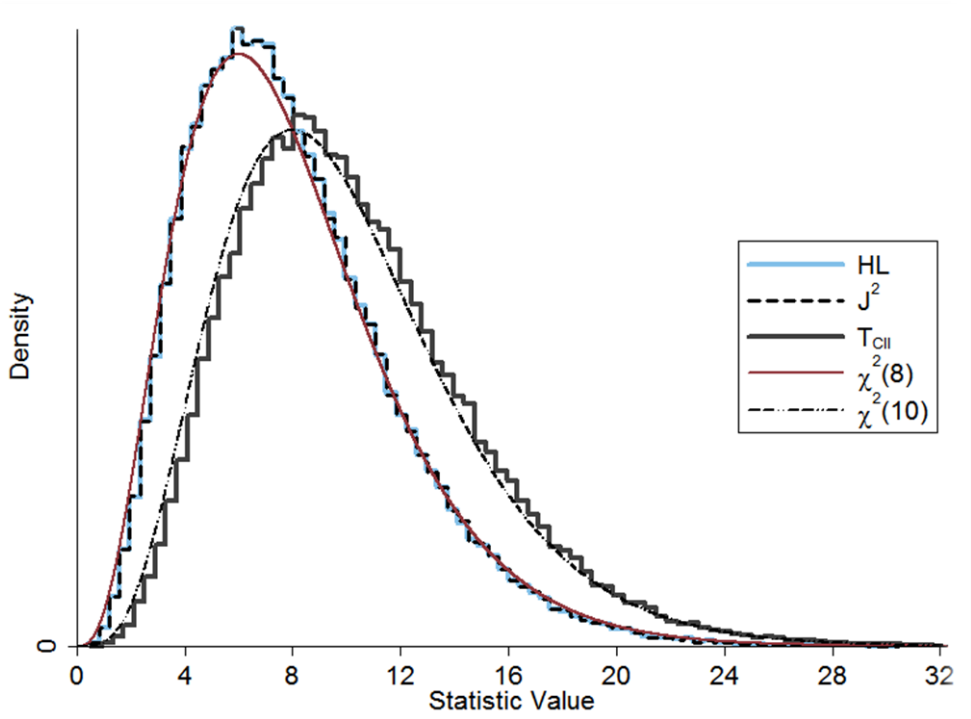

Figure 5.12 Histogram of 100,000 replications of setting 21 using the complementary log- log link. (η=-1.5+0.1x_1+0.1x_2, +0.1 x_1x_2, x_1~U(-1,1), x_2~N(0,1.5), and n=500) The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

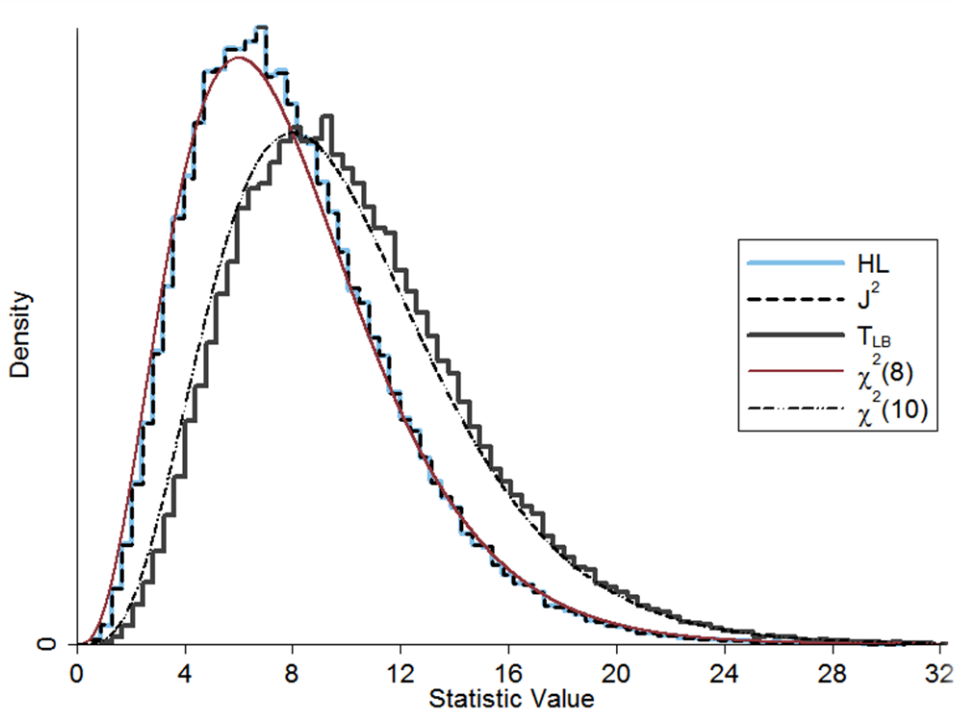

Figure 5.13 Histogram of 100,000 replications of setting 21 using the log link. (η=-1.4+0.1x_1+0.1x_2, +0.1 x_1x_2, x_1~U(-1,1), x_2~N(0,1.5), and n=500) The probability density function curves for $\chi^2(8)$ and $\chi^2(10)$ are included for comparison.

Table 5.4 Summary statistics, rejection per cent, and Kolmogorov-Smirnov test results for *HL*, J^2, and T_G. Simulations were performed using settings 3, 10 and 21 of Table 1 (n=500 and r=100,000) using the deciles-of-risk grouping method with G=10 groups. *HL* and J^2 were compared to critical values of $\chi^2(G\text{-}2)$ and T_G to $\chi^2(G)$ when α=0.10, α=0.05, and α=0.01.

							Percent above			Kolmogorov-Smirnov†
Setting	Link	Statistic	Mean	Std Dev	Skewness	Kurtosis	90th %-ile	95th %-ile	99th %-ile	p-value (D)
3	Probit	*HL*	8.03	3.93	0.98	4.43	9.80	4.78	0.93	0.000 (0.0113)
		J^2	8.06	3.94	0.98	4.43	9.96	4.89	0.96	0.000 (0.0133)
		T_{Pr}	10.05	4.39	0.87	4.13	9.84	4.76	0.93	0.000 (0.0118)
	Log-log	*HL*	8.01	3.94	1.10	5.43	9.73	4.67	0.92	0.000 (0.0087)
		J^2	8.05	3.96	1.10	5.42	9.98	4.82	0.95	0.000 (0.0119)
		T_{LL}	10.01	4.43	0.99	4.96	9.65	4.77	0.95	0.000 (0.0071)
	Complementary log-log	*HL*	8.02	3.94	1.05	4.82	9.66	4.82	1.00	0.000 (0.0112)
		J^2	8.06	3.96	1.05	4.81	9.84	4.93	1.04	0.000 (0.0133)
		T_{Cll}	10.03	4.41	0.92	4.37	9.68	4.82	0.98	0.000 (0.0100)
	Log	*HL*	8.17	4.01	1.31	6.55	10.12	5.21	1.26	0.000 (0.0281)
		J^2	8.21	4.03	1.30	6.51	10.34	5.34	1.30	0.000 (0.0308)
		T_{LB}	9.94	4.45	1.11	5.53	9.47	4.86	1.08	0.000 (0.0137)
10	Probit	*HL*	7.37	3.77	1.04	4.70	7.28	3.49	0.61	0.000 (0.066)
		J^2	7.38	3.78	1.04	4.70	7.33	3.52	0.62	0.000 (0.065)
		T_{Pr}	10.07	4.41	0.87	4.18	10.00	4.84	0.91	0.000 (0.0112)
	Log-log	*HL*	7.39	3.79	1.02	4.61	7.33	3.49	0.63	0.000 (0.0637)
		J^2	7.40	3.80	1.02	4.61	7.37	3.52	0.64	0.000 (0.0625)
		T_{LL}	10.06	4.43	0.87	4.28	10.00	4.90	0.91	0.000 (0.0116)

* The mean was significantly different from that of $\chi^2(8)$ at α=0.05 if outside of the interval (7.975, 8.025). Similarly for $\chi^2(10)$ if outside of (9.972,10.028).

Table 5.4 Summary statistics, rejection per cent, and Kolmogorov-Smirnov test results for *HL*, J^2, and T_G. (cont.)

Setting	Link	Statistic	Mean*	Std Dev	Skewness	Kurtosis	Percent above ** 90th %-ile	95th %-ile	99th %-ile	Kolmogorov-Smirnov† p-value (D)
10	Complementary log-log	*HL*	7.39	3.79	1.02	4.58	7.42	3.47	0.63	0.000 (0.0635)
		J^2	7.40	3.79	1.02	4.58	7.47	3.50	0.63	0.000 (0.0625)
		T_{Cll}	10.07	4.44	1.13	12.34	9.96	4.86	0.96	0.000 (0.0107)
	Log	*HL*	7.37	3.74	0.99	4.45	7.24	3.38	0.54	0.000 (0.0656)
		J^2	7.39	3.75	0.99	4.45	7.32	3.43	0.55	0.000 (0.0635)
		T_{LB}	10.04	4.40	0.85	4.03	9.92	4.86	0.90	0.000 (0.0092)
21	Probit	*HL*	7.79	3.85	0.98	4.38	8.85	4.23	0.75	0.000 (0.0220)
		J^2	7.80	3.86	0.98	4.38	8.89	4.25	0.76	0.000 (0.0211)
		T_{Pr}	10.12	4.47	0.87	4.16	10.40	5.18	1.01	0.000 (0.0128)
	Log-log	*HL*	7.81	3.85	0.99	4.50	8.72	4.23	0.79	0.000 (0.0194)
		J^2	7.82	3.86	0.99	4.50	8.76	4.27	0.80	0.000 (0.0186)
		T_{LL}	10.13	4.48	0.87	4.12	10.39	5.24	1.04	0.000 (0.0186)
	Complementary log-log	*HL*	7.82	3.85	0.98	4.46	8.85	4.22	0.73	0.000 (0.0178)
		J^2	7.82	3.85	0.98	4.46	8.87	4.24	0.74	0.000 (0.0173)
		T_{Cll}	10.09	4.44	0.88	4.22	10.11	4.93	0.97	0.000 (0.0132)
	Log	*HL*	7.87	3.83	0.98	4.47	8.80	4.21	0.76	0.000 (0.0155)
		J^2	7.88	3.83	0.98	4.47	8.84	4.24	0.77	0.000 (0.0149)
		T_{LB}	10.13	4.41	0.86	4.09	10.14	5.00	0.93	0.000 (0.0188)

** The 95% confidence intervals at the α=0.01, α=0.05, and α=0.1 levels were (0.938,1.062), (4.865,5.135) and (9.814, 10.186) respectively.

† H0: No difference between the distribution of the statistic and χ2(df); HA: The distribution of the statistic differs from χ2(df).

most affected by the inclusion of a dichotomous covariate, or the addition of an interaction or quadratic term in the model. The simulations were repeated with only 100 observations under settings 1, 6, 9, 14, and 20. There was no clear pattern when compared to the rates observed when the number of observations was 500. In some cases the rejection percentages were higher while in others they were lower. More than half of the time, the change in $T_{\mathcal{G}}$ went from outside the expected range of rejection percentages under $n = 100$, to within the expected range under $n = 500$. This occurred much less often for the other two statistics. On average, in the null settings the statistics agreed on whether or not to reject the null hypothesis 97% of the time, regardless of the number of observations or the link specified. When there was disagreement between the statistics, in most cases HL and J^2 agreed with each other and disagreed with $T_{\mathcal{G}}$.

5.8.2.2.a Probit

The means and rejection percentages of T_{Pr} were near or within sampling variation in most settings. There were a few exceptions, mainly when the sample size was $n = 100$. The mean and rejection percentages were low in setting 14 ($n = 100$). The rejection percentages were also low in settings 9 ($n = 100$) and 22 ($n = 500$). The mean was high in settings 6, 9, and 20 when $n = 100$, and high in settings 21 and 23 when $n = 500$. The means and rejection percentages of HL and J^2 were lower than expected in over half of the cases studied. In the univariate settings, the rejection percentages of HL and J^2 were very close to or within sampling variation. However, in the settings with more than one covariate, their rejection percentages were often conservative, and in some cases much lower than expected. The lowest rejection percentages observed were among settings where the model contained a dichotomous, a quadratic, or an interaction term.

5.8.2.2.b Log-log

The means and rejection percentages of T_{LL} were again near or within sampling variation in most settings, with some exceptions. The mean and rejection percentages of setting 14 $(n = 100)$ were low. The rejection percentages of settings 1, 14, and 20 where $n = 100$, and 13

and 14 where $n = 500$ were also low. The means of settings 6, 9, 20, 21 ($n = 100$) and 11, 20, 21 ($n = 500$) were high. The mean and especially the standard deviation were also high in setting 11 ($n = 500$). This was the result of seven extreme values of T_{LL}, which were not observed in the other two statistics. When these were removed, the mean of T_{LL} was 10.07, the standard deviation was 4.41, and the rejection per cent was 4.91, while the summary statistics of HL and J^2 were essentially unchanged. The means, standard deviations, and rejection percentages of HL and J^2 were lower than expected in most cases when the settings contained more than one covariate. Very low rejection percentages were observed in some cases when the model contained a dichotomous, a quadratic, or an interaction term. The means were higher than expected in settings 4 and 6 ($n = 100$).

5.8.2.2.c Complementary Log-log

The means and rejection percentages of T_{Cll} were near or within sampling variation in most settings. Exceptions were observed in settings 4 and 14 ($n = 100$) when the rejection percentage was low. Also, the means of settings 1, 6, 9, and 20 ($n = 100$), as well as settings 21 and 22 ($n = 500$) were higher than expected. In setting 11 the standard deviation was higher than expected. In this case there were 6 extreme values of T_{Cll}, with the largest being 377. This was not the case for HL and J^2. When the 6 extreme values were removed, the mean of T_{Cll} was 10.04, the standard variation was 4.41, and the rejection per cent was 4.7. The means of HL and J^2 were higher than expected in a few of the univariate settings and in some of the complex models containing only continuous covariates (settings 1, 6, under $n = 100$ only, and settings 4, 5, 18). Over half of the settings had means, standard deviations, and rejection percentages that were lower than expected, and in some cases they were very low. The lowest rejection percentages were observed when the model contained a dichotomous, a quadratic, or an interaction term.

5.8.2.2.d Log

Both the means and rejection percentages of T_{LB} were near or within sampling variation in most settings. Exceptions were observed in settings 9, 14, and 20 ($n = 100$), as well as 19 ($n = 500$), when rejection percentages were low. The means of settings 6, 9, and 20 ($n = 100$), as well as 21 ($n = 500$), were high. The means of HL and J^2 were higher than expected in most of the univariate settings, while they ranged from higher than expected to much lower than expected in the settings with more complex models. In the univariate settings, the rejection percentages of HL and J^2 were generally within sampling variation. An exception was in setting 1 ($n = 100$, $n = 500$), when the rejection percentages were higher than expected. In this setting, the maximum of the ranges of the predicted probabilities were near 1. In the settings with more complex models, the rejection percentages of HL and J^2 were either within sampling variation or lower than expected, in some cases much lower. The lowest rejection percentages were observed when the model contained a dichotomous, a quadratic, or an interaction term.

5.8.2.3 Power

Simulations were performed to compare the power of HL, J^2, and $T_{\mathcal{G}}$ to detect a departure from a true underlying model under each of the link functions studied. The number of replications run was $r = 10,000$, using some of the settings listed in Table 5.2(a-e) and those in Table 5.3. All settings were run with 500 generated observations ($n = 500$) per dataset. In addition, settings 6, 14, and 20 were also run with 100 observations ($n = 100$) per dataset. The rejection percentage of each statistic, when the linear predictor specified is incorrect, are reported in Table 5.6. The rejection percentages for each statistic when a non-canonical model was incorrectly fit to data generated from logistic settings are reported in Table 5.7.

Overall, the power to detect the departure of a fitted model from that of the true underlying model was similar for all of the statistics. The exception to this was under the log link, when $T_{\mathcal{G}}$ had more power than HL and J^2 to detect an incorrectly fitted log binomial model to logistic data in most settings. The power to detect an incorrectly specified link function when the

Table 5.5(a-d) Simulated null rejection per cent (n=500, r=100,000, α=0.05).

Table 5.5a. Probit Link

		Mean			Standard Deviation			Emperical Rejection Rate (%) †		
Setting	n	HL	J^2	T_{Pr}	HL	J^2	T_{Pr}	HL	J^2	T_{Pr}
1	100	7.86	7.92	9.91	4.41	4.43	4.81	4.61	4.83	4.72
1	500	7.99	8.05	10.05	3.95	3.97	4.42	4.59	4.72	4.99
2	500	7.97	8.02	9.99	3.95	3.97	4.46	4.78	4.94	4.96
3	500	8.03	8.06	10.07	3.91	3.92	4.40	4.77	4.89	4.86
4	500	8.07	8.10	10.02	3.89	3.91	4.47	4.64	4.72	4.99
5	500	8.07	8.10	10.02	4.00	4.01	4.51	5.34	**5.44**	5.01
6	100	8.12	8.17	10.21	3.75	3.77	4.47	**4.44**	**4.50**	5.01
6	500	7.92	7.96	9.96	3.93	3.95	4.48	4.77	4.90	4.87
7	500	7.97	7.98	10.07	3.84	3.84	4.35	**4.34**	**4.38**	4.81
8	500	7.69	7.70	10.06	3.83	3.84	4.45	**4.00**	**4.03**	5.16
9	100	7.92	7.95	10.24	3.76	3.77	4.25	**3.91**	**4.00**	**4.30**
9	500	7.74	7.76	10.04	3.86	3.87	4.44	**4.22**	**4.25**	4.97
10	500	7.28	7.30	10.07	3.74	3.75	4.42	**3.47**	**3.47**	4.97
11	500	7.17	7.21	10.04	3.70	3.72	4.42	**3.09**	**3.17**	4.80
12	500	7.73	7.73	9.99	3.87	3.88	4.44	**4.17**	**4.18**	4.94
13	500	7.11	7.15	10.03	3.66	3.68	4.39	**2.67**	**2.74**	4.73
14	100	7.21	7.26	9.84	3.57	3.59	4.14	**2.86**	**2.94**	**3.65**
14	500	7.38	7.43	10.14	3.77	3.79	4.46	**3.23**	**3.35**	5.10
15	500	7.64	7.68	10.05	3.81	3.83	4.43	**3.51**	**3.64**	4.67
16	500	7.92	7.97	10.09	3.87	3.89	4.40	4.62	4.81	4.94
17	500	7.94	7.95	9.96	3.87	3.88	4.37	**4.56**	4.58	4.68
18	500	7.98	7.99	10.06	3.86	3.86	4.39	**4.48**	**4.54**	4.91
19	500	7.19	7.19	10.06	3.68	3.68	4.39	**2.97**	**2.98**	4.75
20	100	6.97	7.02	10.38	3.37	3.39	4.20	**2.08**	**2.14**	4.84
20	500	6.82	6.86	10.09	3.59	3.61	4.45	**2.30**	**2.41**	5.02
21	500	7.84	7.84	10.15	3.89	3.90	4.50	**4.28**	**4.30**	5.25
22	500	7.34	7.39	10.06	3.58	3.61	4.33	**3.00**	**3.08**	**4.53**
23	500	7.79	7.84	10.14	3.75	3.78	4.35	**3.83**	**3.96**	4.74
24	500	7.88	7.94	10.09	3.88	3.91	4.43	**4.37**	**4.51**	5.05

† Rejection percentages were significantly different from 5% at α=0.05 level if they fell above 5.43 or below 4.57 (bold).

Table 5.5b. Log-log Link

		Mean			Standard Deviation			Emperical Rejection Rate (%) †		
Setting	n	HL	J^2	T_{LL}	HL	J^2	T_{LL}	HL	J^2	T_{LL}
1	100	8.02	8.07	10.04	3.84	3.86	4.29	**4.30**	**4.47**	**4.13**
1	500	8.04	8.09	10.10	3.92	3.94	4.40	4.96	5.16	5.02
2	500	8.03	8.07	10.06	3.90	3.92	4.37	4.87	4.97	4.71
3	500	7.99	8.04	10.00	3.94	3.96	4.42	**4.56**	4.67	4.75
4	500	8.10	8.13	10.10	3.96	3.97	4.48	4.81	4.90	5.03
5	500	8.05	8.07	9.98	3.88	3.90	4.36	4.62	4.63	4.58
6	100	8.17	8.21	10.23	3.77	3.79	4.33	**4.20**	**4.29**	4.83
6	500	8.03	8.06	10.05	3.98	3.99	4.48	4.73	4.80	4.86
7	500	7.99	8.00	10.07	3.96	3.97	4.49	4.66	4.68	4.94
8	500	7.85	7.87	10.06	3.86	3.87	4.44	**4.21**	**4.24**	4.99
9	100	7.97	8.00	10.31	3.77	3.78	4.26	**4.11**	**4.20**	**4.56**
9	500	7.89	7.91	9.99	3.90	3.91	4.42	4.62	4.65	5.05
10	500	7.37	7.38	10.03	3.78	3.78	4.42	**3.60**	**3.62**	4.83
11	500	7.22	7.25	10.33	3.71	3.73	19.71	**3.08**	**3.19**	4.98
12	500	7.88	7.89	10.00	3.94	3.95	4.45	4.59	4.60	4.73
13	500	7.10	7.14	9.99	3.61	3.63	4.37	**2.75**	**2.83**	**4.51**
14	100	7.21	7.26	9.86	3.53	3.55	4.15	**2.64**	**2.74**	**3.73**
14	500	7.31	7.35	10.12	3.75	3.77	4.45	**3.16**	**3.28**	4.93
15	500	7.74	7.79	10.11	3.88	3.90	4.44	**4.33**	**4.43**	4.98
16	500	7.81	7.86	9.97	3.86	3.89	4.38	**4.48**	4.65	4.80
17	500	7.97	7.98	10.04	3.97	3.98	4.48	4.96	4.99	5.16
18	500	7.95	7.97	10.03	3.94	3.95	4.49	4.76	4.80	5.08
19	500	7.22	7.22	10.10	3.71	3.71	4.45	**3.04**	**3.04**	5.13
20	100	6.94	6.98	10.32	3.31	3.34	4.13	**1.89**	**1.98**	**4.30**
20	500	6.85	6.89	10.14	3.57	3.59	4.44	**2.27**	**2.29**	5.12
21	500	7.86	7.87	10.16	3.89	3.90	4.50	**4.47**	**4.54**	5.27
22	500	7.25	7.29	10.03	3.63	3.65	4.43	**3.00**	**3.13**	4.88
23	500	7.63	7.67	10.06	3.77	3.79	4.45	**3.61**	**3.72**	4.98
24	500	7.71	7.76	10.04	3.78	3.80	4.41	**4.11**	**4.14**	5.06

† Rejection percentages were significantly different from 5% at α=0.05 if they fell above 5.43 or below 4.57 (bold).

Table 5.5c. Complementary log-log Link

		Mean			Standard Deviation			Emperical Rejection Rate (%) †		
Setting	n	*HL*	J^2	T_{CII}	*HL*	J^2	T_{CII}	*HL*	J^2	T_{CII}
1	100	8.13	8.18	10.17	3.87	3.89	4.27	4.66	4.80	4.82
1	500	8.04	8.08	10.06	3.91	3.93	4.40	4.91	5.10	4.80
2	500	8.03	8.04	10.04	3.92	3.92	4.37	4.86	4.93	4.73
3	500	8.03	8.07	10.05	3.95	3.97	4.44	4.81	5.01	4.78
4	500	8.17	8.19	10.08	3.84	3.84	4.25	4.72	4.74	**4.35**
5	500	8.10	8.11	10.04	3.97	3.98	4.45	5.00	5.06	5.04
6	100	8.18	8.21	10.34	3.89	3.90	4.46	4.65	4.82	5.39
6	500	8.08	8.10	10.11	4.01	4.02	4.53	5.20	5.31	5.17
7	500	8.01	8.02	10.11	3.93	3.93	4.40	4.66	4.68	4.71
8	500	7.82	7.84	10.03	3.85	3.86	4.41	**4.41**	**4.50**	5.22
9	100	7.87	7.89	10.32	3.78	3.79	4.31	**3.83**	**3.92**	**4.56**
9	500	7.92	7.94	10.03	3.90	3.91	4.46	**4.34**	**4.37**	5.02
10	500	7.35	7.36	10.01	3.74	3.74	4.42	**3.46**	**3.51**	4.86
11	500	7.17	7.22	10.12	3.70	3.72	6.20	**2.96**	**3.07**	4.76
12	500	7.82	7.83	9.99	3.92	3.92	4.48	4.59	4.60	5.06
13	500	7.12	7.16	10.02	3.67	3.69	4.37	**2.84**	**2.92**	4.77
14	100	7.36	7.41	10.04	3.52	3.54	4.11	**2.71**	**2.80**	**3.79**
14	500	7.37	7.41	10.11	3.74	3.76	4.44	**3.50**	**3.61**	5.34
15	500	7.84	7.89	10.13	3.92	3.94	4.47	**4.43**	**4.54**	5.24
16	500	7.96	8.01	10.06	3.93	3.95	4.45	**4.55**	4.64	5.03
17	500	7.97	7.98	10.04	3.95	3.96	4.44	4.82	4.84	4.87
18	500	8.10	8.12	10.09	3.97	3.98	4.49	5.07	5.17	5.30
19	500	7.19	7.20	10.04	3.61	3.61	4.34	**2.70**	**2.72**	**4.54**
20	100	7.12	7.16	10.42	3.40	3.42	4.22	**2.06**	**2.11**	4.74
20	500	6.89	6.93	10.06	3.57	3.58	4.38	**2.40**	**2.47**	4.72
21	500	7.87	7.88	10.14	3.87	3.87	4.43	**4.47**	**4.49**	5.05
22	500	7.46	7.49	10.16	3.75	3.76	4.48	**3.58**	**3.64**	5.30
23	500	7.68	7.72	10.02	3.72	3.74	4.35	**3.82**	**3.95**	**4.55**
24	500	7.89	7.92	10.05	3.82	3.84	4.37	**4.43**	**4.50**	4.83

† Rejection percentages were significantly different from 5% at α=0.05 if they fell above 5.43 or below 4.57 (bold).

Table 5.5d. Log Link

		Mean			Standard Deviation			Emperical Rejection Rate (%) †		
Setting	n	HL	J^2	T_{LB}	HL	J^2	T_{LB}	HL	J^2	T_{LB}
1	100	8.03	8.08	9.88	4.28	4.29	4.63	**5.48**	**5.53**	5.01
1	500	8.23	8.29	9.81	4.07	4.09	4.46	**5.68**	**5.84**	4.77
2	500	8.04	8.05	10.04	3.99	3.99	4.44	4.80	4.83	4.71
3	500	8.14	8.19	9.89	4.02	4.04	4.43	5.27	5.43	4.80
4	500	8.31	8.33	10.12	4.03	4.04	4.43	5.31	5.39	4.91
5	500	8.14	8.15	9.98	3.97	3.97	4.40	5.30	5.33	4.96
6	100	8.11	8.13	10.24	3.75	3.76	4.29	**4.27**	**4.40**	4.64
6	500	8.07	8.10	9.98	3.89	3.90	4.38	4.68	4.75	4.60
7	500	8.03	8.04	10.11	3.86	3.87	4.40	4.68	4.70	4.94
8	500	8.07	8.10	9.99	3.88	3.89	4.41	4.81	4.94	4.64
9	100	7.98	8.00	10.23	3.76	3.77	4.24	**3.95**	**4.02**	**4.12**
9	500	8.03	8.06	10.04	3.89	3.90	4.42	4.83	4.89	4.95
10	500	7.34	7.36	10.02	3.71	3.72	4.39	**3.40**	**3.42**	4.84
11	500	7.11	7.20	10.03	3.69	3.73	4.44	**2.97**	**3.16**	4.85
12	500	7.84	7.85	10.00	3.93	3.93	4.47	**4.45**	**4.46**	4.82
13	500	7.27	7.31	10.01	3.73	3.75	4.41	**3.39**	**3.47**	4.74
14	100	7.34	7.38	10.00	3.53	3.55	4.17	**2.52**	**2.58**	**3.82**
14	500	7.48	7.52	10.07	3.77	3.79	4.45	**3.49**	**3.60**	4.97
15	500	7.97	8.01	10.11	3.92	3.93	4.47	4.65	4.82	5.04
16	500	8.16	8.21	10.05	3.96	3.98	4.43	4.92	5.07	4.71
17	500	8.13	8.14	10.00	3.96	3.97	4.43	5.33	5.36	5.08
18	500	8.21	8.23	10.14	3.98	3.99	4.46	5.05	5.09	4.82
19	500	7.18	7.19	10.04	3.65	3.65	4.38	**2.84**	**2.84**	**4.55**
20	100	7.10	7.13	10.31	3.34	3.35	4.13	**2.03**	**2.06**	**4.30**
20	500	7.07	7.10	10.01	3.58	3.60	4.41	**2.62**	**2.69**	4.78
21	500	7.90	7.91	10.17	3.90	3.90	4.47	**4.33**	**4.37**	5.21
22	500	7.57	7.61	10.06	3.66	3.67	4.35	**3.45**	**3.55**	4.70
23	500	7.82	7.86	10.02	3.81	3.83	4.38	**4.27**	**4.36**	4.94
24	500	8.06	8.10	9.99	3.87	3.88	4.37	4.68	4.79	4.69

† Rejection percentages were significantly different from 5% at α=0.05 if they fell above 5.43 or below 4.57 (bold).

underlying model was logistic was highest when the incorrectly specified link was either log or log-log. On average, in the power settings the statistics agreed on whether or not to reject the null hypothesis when an incorrect link was specified between 91% and 96% of the time. When a term was incorrectly omitted from the linear predictor, they agreed between 95% and 97% of occasions if $n = 500$, and between 93% and 96% of occasions when $n = 100$. When there was disagreement between the statistics, most of the time HL and J^2 agreed with each other and disagreed with $T_{\mathcal{G}}$.

5.8.2.3.a Probit

Under the probit model, the power to detect either an incorrectly specified probit link function or the linear predictor with an omitted term was similar for all the three statistics. The power to detect the omission of a quadratic term ranged from low to high, with power increasing as the lack of linearity became more pronounced or when the number of observations increased from 100 to 500. The power to detect an omitted interaction term was low to moderate. There was low power to detect the omission of a continuous covariate. In general, the statistics had more power to detect the omission of terms from the linear predictor than to detect an incorrectly specified link function. All had very little power to identify an incorrect probit model fitted to logistic data. This was not surprising, due to the similarities between the logistic and probit curves. On average, when an incorrect probit link was specified, the statistics agreed on whether or not to reject the null hypothesis 96% of the time. When the linear predictor of the probit model had a term omitted, they agreed 97% of the time when $n = 500$ and 96% of the time when $n = 100$. When there was disagreement between the statistics, usually HL and J^2 agreed with each other and disagreed with T_{Pr}.

5.8.2.3.b Log-log

Under the log-log model, the power to detect either an incorrectly specified log-log link function or the linear predictor with an omitted term was similar for all the three statistics. The power to detect the omission of a quadratic term ranged from low to high, with power

increasing as the lack of linearity became more pronounced or when the number of observations increased from 100 to 500. The power to detect an omitted interaction term was low to high. There was low power to detect the omission of a continuous covariate. In general, the statistics had more power to detect the omission of terms from the linear predictor than to detect an incorrectly specified link function. All had very little power to detect an incorrect log-log model fitted to logistic data. On average, when an incorrect log-log link was specified, the statistics agreed on whether or not to reject the null hypothesis 94% of the time. When the linear predictor of the log-log model had a term omitted, they agreed 97% of the time when $n = 500$ and 93% of the time when $n = 100$. When there was disagreement between the statistics, usually HL and J^2 agreed with each other and disagreed with T_{LL}.

5.8.2.3.c Complementary Log-log

Under the complementary log-log model, the power to detect either an incorrectly specified link function or a linear predictor with an omitted term was similar for all the three statistics. All three of the statistics had generally more power to detect the omission of terms in the linear predictor than to detect an incorrectly specified link function. The power to detect the omission of a quadratic term ranged from low to high, with power increasing as the lack of linearity became more pronounced or when the number of observations increased from 100 to 500. The power to detect an omitted interaction term was low to moderate. There was low power to detect the omission of a continuous covariate. The power to detect an incorrect complementary log-log model fitted to logistic data was very low, similar to that under the probit settings. On average, when a complementary log-log link was incorrectly specified, the statistics agreed on whether or not to reject the null hypothesis 96% of the time. When the linear predictor of the complementary log-log model had a term omitted, they agreed 96% of the time when $n = 500$ and 95% of the time when $n = 100$. When there was disagreement between the statistics, usually HL and J^2 agreed with each other and disagreed with T_{Cll}.

5.8.2.3.d Log

The power to detect an incorrect model when a term was omitted from the linear predictor was similarly low among the three statistics; however, when an incorrectly specified log link function was fitted to logistic data, T_{LB} had more power than either HL or J^2 to detect the incorrect model in most settings. The largest differences in power were observed in settings with either a dichotomous term, a quadratic term, or a strongly influential interaction term, with the greatest difference in the percentage rejected observed (about 14%) in setting 27, when the linear predictor contained a continuous and a dichotomous covariate. The power to detect the omission of a quadratic term ranged from low to high, with power increasing as the lack of linearity became more pronounced. The power to detect an omitted interaction term was low to moderate. There was low power to detect the omission of a continuous covariate. On average, when an incorrect log link was specified, the statistics agreed on whether or not to reject the null hypothesis 91% of the time. When the linear predictor of the log binomial model had a term omitted, they agreed 95% of the time when $n = 500$ and 96% of the time when $n = 100$. When there was disagreement between the statistics, usually HL and J^2 agreed with each other and disagreed with T_{LB}.

5.9 Examples

In two examples, we fit logistic, probit, log-log, complementary log-log, and log binomial regression models to data from the Low Birth Weight study described in *Applied Logistic Regression* (2000). The models relate an outcome of a low birth weight among babies to variables that are measures of their mother's behaviour and physical characteristics. In both examples we chose covariates with significant Wald test statistics $(\alpha = 0.05)$ for at least one of the link functions. We then compared the GOF statistics for the models under all of the links.

The first group of models, each fitted with a different link function, contain race (white, black, other) and the weight of the mother at her last menstrual period as predictors. These variables were considered medically important (Hosmer, et al. 2000). In this case, all of the covariates had significant Wald test statistics at the 5% level when the logistic model was fit to the data,

Table 5.6 Empirical rejection per cent (α=0.05) when a model with a term omitted from the linear predictor was fitted to data generated from a model with all terms.

		Probit			Log-Log			Complementary log-log			Log		
Setting	n	HL	J^2	T_{Pr}	HL	J^2	T_{Pr}	HL	J^2	T_{Pr}	HL	J^2	T_{LB}
6	100	4.5	4.6	4.4	4.2	4.3	4.1	4.8	4.9	4.6	4.2	4.2	4.0
6	500	5.3	5.3	5.4	4.9	4.9	4.8	5.3	5.3	5.0	4.7	4.7	4.8
7	500	5.2	5.2	5.1	5.1	5.1	4.8	5.0	5.0	4.9	5.0	5.0	5.0
13	500	11.9	12.1	11.6	48.9	49.1	49.7	4.8	4.9	4.9	16.9	17.3	19.5
14	100	32.4	32.6	31.2	26.4	26.6	27.2	25.8	26.0	24.5	18.2	18.4	17.2
14	500	88.5	88.6	87.9	89.9	90.0	89.7	66.7	66.8	65.5	50.8	51.1	50.3
15	500	99.0	99.0	98.9	99.7	99.7	99.7	97.2	97.2	97.0	91.2	91.2	92.0
16	500	100	100	100	100	100	100	100	100	100	100	100	100
19	500	7.3	7.3	7.9	5.5	5.5	6.0	5.2	5.2	5.8	4.6	4.6	5.5
20	100	24.3	24.5	26.1	29.9	30.2	32.9	15.9	16.1	17.0	9.4	9.5	9.9
20	500	51.6	51.8	54.4	98.8	98.8	98.7	62.3	62.5	61.6	33.5	33.7	32.4
21	500	6.9	7.0	7.9	7.2	7.2	7.8	5.7	5.7	6.1	5.9	5.9	6.5
22	500	68.7	69.0	68.3	67.1	67.4	67.9	38.6	38.8	39.0	32.8	33.0	33.9
23	500	97.7	97.7	97.9	96.6	96.7	96.7	76.2	76.4	78.6	65.3	65.4	69.3
24	500	99.9	99.9	100	99.9	99.9	99.9	91.1	91.2	94.1	88.0	88.0	92.4

Table 5.7 Empirical rejection per cent (α=0.05) when a model with an incorrectly specified link function was fitted to data generated from an underlying logistic model.

	Probit			Log-log			Complementary log-log			Log		
Setting	*HL*	*J2*	*TLB*	*HL*	*J2*	*TLB*	*HL*	*J2*	*TLB*	*HL*	*J2*	*TLB*
25	5.5	5.6	5.8	17.0	17.2	16.7	7.5	7.6	7.7	30.4	30.8	32.7
26	7.7	7.7	7.5	16.0	16.0	17.6	4.3	4.4	4.6	4.4	4.5	4.7
27	8.2	8.4	8.7	23.6	23.8	24.6	3.9	4.0	5.6	7.8	8.1	21.8
28	8.8	8.9	9.0	30.9	31.1	33.6	5.2	5.4	6.3	20.9	21.3	28.1
29	7.6	7.7	8.2	26.6	26.8	28.8	5.8	5.9	6.5	20.3	20.6	26.8
30	7.8	7.9	8.6	33.6	33.8	35.7	6.7	6.9	8.0	30.5	31.1	42.1
31	2.9	3.0	5.2	2.8	2.8	4.7	3.1	3.2	5.2	3.6	3.6	5.6
32	4.0	4.0	5.6	6.9	7.1	9.3	3.7	3.7	6.2	6.9	7.1	11.5
33	4.9	5.0	5.7	10.3	10.5	11.0	3.8	3.9	5.9	6.7	6.9	14.8
34	5.8	5.9	5.8	13.6	13.9	13.4	4.0	4.2	6.1	9.1	9.3	18.6
35	6.2	6.3	6.6	15.0	15.2	15.4	4.1	4.2	6.0	9.5	9.9	21.8
36	4.7	4.7	4.7	5.6	5.7	5.9	5.3	5.4	5.7	8.2	8.2	10.3
37	5.0	5.0	5.5	5.1	5.2	5.7	5.3	5.3	5.8	6.1	6.3	7.9
38	3.1	3.1	4.9	2.9	2.9	4.8	3.1	3.1	4.8	3.2	3.2	5.0
39	2.5	2.5	4.7	2.6	2.6	5.2	2.5	2.6	4.7	2.7	2.7	5.7
40	2.6	2.7	4.9	3.7	3.8	7.2	2.7	2.9	6.4	7.4	7.6	18.2

except for one of the race categories which was kept in the model because it was one of the design variables.

For the second set of models, the covariates selected include the history of hypertension, smoking status during pregnancy, race (white, black or other), and the weight, raised to the third power, of the mother at her last menstrual period (i.e. lwt^3). In this case, the covariates were not chosen based on medical theory, but based only on Wald test significance levels $(\alpha = 0.05)$ obtained when a log model was fit to the data.

The HL, J^2, and $T_{\mathcal{G}}$ statistics were calculated to assess the fit of these models to the observed data, and are reported in Table 5.8. The Low Birth Weight dataset contains 189 observations. Groups were formed by ordering the predicted probabilities, and placing 19 in each of the first nine groups and 18 in the tenth. Ties were placed into the same group. When they occurred on group boundaries, the placement of the ties into the upper or lower group was consistent among the different links.

All three statistics gave similar results in both examples. In the first set of models, the p-values of the three statistics indicate that all of the models give a reasonable fit to the data, with the complementary log-log model having the closest agreement between the expected and observed data values. In the second set of models, all three statistics indicated that both the logit and the log-log models fit the data poorly, with p-values less than or near 0.05. All three statistics indicate that either the complementary log-log or the log models gave the closest fit.

Table 5.8 Values of *HL*, J^2, and $T_{\mathcal{G}}$, along with their associated p-values, calculated for models using data from the Low Birth Weight Study described in Applied Logistic Regression (Hosmer, et al. 2000).

Model	Link	*HL*	p-value	J^2	p-value	*T*	p-value
1	logit	7.60	0.47	7.61	0.47	8.20	0.51
	probit	7.59	0.47	7.89	0.44	8.53	0.58
	log-log	8.25	0.41	8.25	0.41	8.70	0.56
	complementary log-log	5.28	0.73	5.29	0.73	7.45	0.68
	log	9.78	0.28	9.79	0.28	10.18	0.43
2	logit	16.77	0.03	16.92	0.03	17.93	0.04
	probit	12.68	0.12	12.79	0.12	14.26	0.16
	log-log	15.36	0.05	15.46	0.05	17.51	0.06
	complementary log-log	10.44	0.24	10.54	0.23	10.96	0.36
	log	9.62	0.29	10.01	0.26	10.31	0.41

5.10 Discussion

The statistic, $T_{\mathcal{G}}$, based on the Tsiatis goodness-of-fit test for logistic regression, is proposed as a goodness-of-fit statistic for GLMs for Bernoulli outcome data with any link function. The distributional properties and the performance of $T_{\mathcal{G}}$ under four non-canonical links (probit, log-log, complementary log-log, and log) was compared to that of two other test statistics developed originally for logistic regression, HL and J^2. The deciles-of-risk method was used to form groups for all three statistics.

The distributions of the statistics under each link function were compared to their postulated distributions. That is, HL and J^2 were both compared to $\chi^2(G-2)$, and $T_{\mathcal{G}}$ to $\chi^2(G)$. In the settings investigated, the distribution of $T_{\mathcal{G}}$ followed $\chi^2(G)$ closely under all four links, but the distributions of HL and J^2 were more dependent on the characteristics of the model. When the model contained only a continuous covariate, the distributions of HL and J^2 were close to $\chi^2(G-2)$, but when the linear predictor contained one continuous covariate and a dichotomous term, or two continuous covariates and an interaction term, the curves tended to be shifted left and to have higher peaks than that of $\chi^2(G-2)$. The Type I error rates reflected the

results of the histograms. The empirical rejection percentages of $T_{\mathcal{G}}$ were within the values expected in nearly all cases examined, but this was not true for HL and J^2. Although their Type I error rates were well maintained overall when the model setting contained only a single continuous covariate, they were generally lower than expected when the model also contained more than one term, especially when one of the terms was a dichotomous, quadratic, or interaction term. The results in the case of the log binomial link were similar to those reported by Blizzard, et al. (2006). Because the results indicate that the distribution of HL or J^2 varies depending on the characteristics of the underlying model, it is unclear how an adjustment to the degrees-of-freedom could be made to improve performance, and thus it is recommended that HL and J^2 be compared to $\chi^2(G-2)$ under non-canonical links, when the deciles-of-risk grouping method is used.

The results of the power simulations indicate that all three statistics had only low to moderate power to detect an incorrectly specified link function when the true underling model was logistic. This is not very surprising, given the potential similarities between the shape of the logistic function and those of some of the non-canonical GLM. For example, the lower tail of the logit, probit, log, and complementary log-log functions can be very similar, as can the upper tail of the logit, probit, and log-log functions. Potential differences are illustrated in Figure 2.2, with the largest difference usually occurring between the upper tails of the log function and the other link functions when the range of the predicted probabilities includes larger values, much greater than 0.5. Under the probit and complementary log-log links, we observed that the power was low. However in some settings, when the link function was log or log-log, the power of all three statistics was moderate. Under the log link, $T_{\mathcal{G}}$ had more power than either HL or J^2 in many of the settings. All of the statistics had very similar power to detect the omission of a term from the linear predictor. The power ranged from very low to very high, with the greatest power observed when the term omitted was a quadratic or interaction term, especially when these terms had greater influence in the model. The power to detect the omission of a continuous covariate was poor in all cases.

The results of the simulation study indicate that $T_{\mathcal{G}}$ offers some advantages over HL or J^2 when evaluating the fit of non-canonical GLM, particularly when the log link function is chosen. Under all of the links studied, $T_{\mathcal{G}}$ maintained the Type I error rate well, regardless of the number or types of terms included in the model. This was not the case for HL and J^2. In addition, our simulation study indicates that under the log link, T_{LB} offers higher power to detect an incorrectly specified log link when the true underlying model is logistic.

Chapter 6 Overall Discussion

6.1 Overview of the Chapter

We begin by giving an overview of the earlier chapters, and then outline the structure of the current chapter. The background material for this work was discussed in Chapters 1 to 3. In Chapter 4, an analysis was made comparing the algebraic forms of three goodness-of-fit statistics, HL, J^2, and T, all of which can be used to assess the fit of logistic regression models with continuous covariates. A simulation study was also presented that investigates the distributional characteristics of these statistics and compares their performances when applied to settings with finite samples. In Chapter 5, a generalized form of the Tsiatis statistic, $T_{\mathcal{G}}$, was introduced. It can be applied to any GLM with Bernoulli outcomes, regardless of link function. We show that when $T_{\mathcal{G}}$ was derived for the canonical logistic model, it reduces to the usual Tsiatis statistic, T. We also derived the terms of $T_{\mathcal{G}}$ for four non-canonical GLM in common use: the probit, log-log, complementary log-log, and log binomial models. Corresponding forms of HL and J^2 are also given. A simulation study was conducted to investigate whether the distribution of the new statistic when applied to a finite sample and when the deciles-of-risk grouping method is used would be approximately $\chi^2(G)$.In addition, the performance of the three statistics when evaluating probit, log-log, complementary log-log, and log binomial models was also studied. Chapter 5 concludes with the presentation of two examples where HL, J^2, and $T_{\mathcal{G}}$ are used to assess the fit of probit, log-log, complementary log-log and log binomial models that were fitted to a real world dataset. In this current chapter we synthesize the results of the earlier chapters, and discuss how the work fits into the broader scope of current research. We discuss what contribution it makes, why this is significant, and the limitations of the work. Finally, we make suggestions for future research.

6.2 Broad View of Research

The relationship between observed Bernoulli outcomes and a set of explanatory covariates can be modelled with a GLM. The most commonly used GLM in this situation is the logistic regression model. One critical step in the evaluation of any GLM is to assess how well it fits the observed data. When the explanatory covariates are discrete and the model estimated frequencies are not small, it is appropriate to use two well-known goodness-of-fit statistics: the deviance and Pearson's chi-squared. However, when the model contains continuous covariates, or covariates that behave as though they were continuous, then the distributional theory behind these statistics is violated, and their use is not appropriate. Other goodness-of-fit statistics must be used for the assessment of the model fit. Several statistics have been proposed that employ artificial grouping methods.

The Hosmer-Lemeshow goodness-of-fit statistic is one of the most widely used statistics for evaluating the fit of logistic regression models containing continuous covariates. Usually, HL is calculated using the deciles-of-risk grouping method. This method forms groups based on the predicted probabilities. Under this method, the group boundaries are determined by referencing the random outcome data; thus, the boundaries are themselves random. This results in a reduction of the degrees of freedom of the distribution of the statistic. Extensive simulations by Hosmer, et al. (1980) confirmed that the asymptotic distribution of HL is approximately $\chi^2(G-2)$.

Two other goodness-of-fit statistics for logistic regression are J^2 (Pigeon, et al. 1999b) and T (Tsiatis 1980). When calculating these two statistics, artificial grouping of the data is usually accomplished through partitioning the covariate space. Because partitioning is performed without reference to the observed data, the boundaries of the partition are not random. Pigeon and Heyse state that the distribution of J^2 does not rely on a particular grouping strategy. They report the asymptotic distribution of J^2 to be approximately $\chi^2(G-1)$, regardless of grouping method, which suggests that applying other methods such as the deciles-of-risk should not affect its distribution. Halteman (1980), in his unpublished thesis, evaluated the distributional

characteristics of T when the deciles-of-risk method was used. He found that the distribution of T was unaffected by random cell boundaries. Through simulations he found that under finite samples, the distribution of T was approximately $\chi^2(G-1)$.

6.3 Need For This Research

This work addresses several questions. One topic that we address is how HL, J^2, and T are related algebraically. Another is whether there is any difference in their performances. Although the performance of HL has been compared to many other omnibus goodness-of-fit statistics (Hosmer, et al. 1997, Kuss 2002), we know of no published simulation studies that assess the performances of HL, J^2, and T under the same grouping method. We applied the deciles-of-risk method to all three statistics in order to isolate any differences in performance attributable to differences in their algebraic forms.

In addition, we sought to verify the reported distributions of the three statistics. Specifically, we checked whether HL had a distribution that was approximately $\chi^2(G-2)$, and sought to confirm that J^2, and T had distributions that were approximately $\chi^2(G-1)$.

Few goodness-of-fit statistics have been developed for non-canonical GLM with Bernoulli outcomes when the models contain continuous covariates. Farrington (1995, 1996) considered goodness-of-fit methods for non-canonical GLM; however, these methods only address the assessment of the fit of models with discrete covariates. Blizzard, et al. (2006, 2007), have considered the distributional characteristics and performance of HL when applied to binary log binomial models. Other statistics they evaluated for assessing the fit of log binomial models included the standard normalized version of the Pearson chi-square (Osius, et al. 1992), the standardized unweighted sum of squares, the Stukel score test (Stukel 1988), the Hjort-Hosmer statistic (Hosmer, et al. 2002), and the le Cressie and van Houwelingen statistic (le Cessie, et al. 1991). Without appropriate test statistics for evaluating the fit of non-canonical GLM, a less than optimal model may be selected. A poorly selected model could affect the interpretation of research results.

6.4 Contribution and Significance of This Research

Our research shows that HL, J^2, and T are closely related algebraically. We observed that the J^2 statistic is simply the T statistic calculated with the covariance matrix (4.3) rather than the conditional covariance matrix (5.14). We showed that $HL \leq J^2$, and also under what conditions J^2 is much larger than HL.

In a simulation study, we examined the distributional properties and performance of HL, J^2, and T when assessing logistic regression models with continuous covariates under the deciles-of-risk grouping method. We found that HL and T followed their reported approximate asymptotic null distributions, $\chi^2(G-2)$ and $\chi^2(G-1)$ respectively, but that J^2 did not. Instead, we found the distribution of J^2 to be much closer to $\chi^2(G-2)$, rather than the reported $\chi^2(G-1)$ (Pigeon, et al. 1999b), when calculated using the deciles-of-risk grouping method.

When HL and J^2 were assumed to follow a $\chi^2(G-2)$ distribution, and T a $\chi^2(G-1)$ distribution, we found that T controlled the Type I error rate ($\alpha = 0.05$) about twice as often as HL and J^2. The null rejection percentages of all three statistics were at or near the values expected when the linear predictor contained only continuous covariates. However, the Type I error rates of HL and J^2, in some settings, were lower than expected when the models contained either a dichotomous, quadratic, or interaction term. Curiously, all three statistics were observed to have Type I error rates that were higher than expected in a series of settings with two continuous covariates and a quadratic term when the sample size was small $(n = 100)$.

All three statistics had similar power to detect an incorrectly specified logistic model. The statistics agreed on whether to reject the null hypothesis about 97% of the time. However, when they did not agree, in most cases HL and J^2 agreed with each other and disagreed with T. When the linear predictor was incorrectly specified, all three statistics had low power to detect the omission of an interaction term and the omission of a quadratic term that had limited departure from linearity. As the departure from linearity of the quadratic term increased, the

power of each of the statistics to detect the omission of the quadratic term also increased. In one series of settings in which the nonlinearity of the model increased, the power became very high. Finally, all three of the statistics generally had low power to detect an incorrectly specified link function, but there were two exceptions. The first was when the true model had asymmetric tails and the sample size was $n = 500$. In this case the power of all three statistics was very high. The second was when the true model was a complementary log-log model, again with $n = 500$. In this case the power of T was higher than the other two statistics, although all three had low power.

The original Tsiatis goodness-of-fit score test for logistic regression is a quadratic form of observed counts minus expected counts. We found, however, that this test statistic could not be applied to non-canonical GLM, as it did not result in this form under non-canonical link functions. By augmenting the original Tsiatis model with an additional term, we generalized the original Tsiatis score test so that when it is applied to any GLM with Bernoulli distributed outcomes, regardless of link function, it will result in a quadratic form of observed counts minus expected counts. We derived both the general formula for the new statistic and the specific terms of the statistic under the probit, log-log, complementary log-log, and log binomial models. Unlike the original Tsiatis statistic, we showed that the conditional covariance matrix used in the calculation of $T_{\mathcal{G}}$ is nonsingular. Thus the calculation of a generalized inverse is not required. This simplifies the calculation of the statistic. This, along with the simplicity of the deciles-of-risk grouping method, may make this test statistic more accessible to researchers who lack skills in advanced mathematics.

We conducted a simulation study comparing the distributional characteristics and performance of HL, J^2, and $T_{\mathcal{G}}$, when assessing the fit of probit, log-log, complementary log-log, and log binomial models with continuous covariates. We postulated, based on our earlier results and the results in section 5.4, that the distributions of HL and J^2 were approximately $\chi^2(G-2)$, while the distribution of $T_{\mathcal{G}}$ was approximately $\chi^2(G)$. Our results, from three simulation runs with very high replication rates, indicate that the distribution of $T_{\mathcal{G}}$ followed $\chi^2(G)$ closely

under all four links, but the distributions of HL and J^2 were more dependent on the characteristics of the model. The distributions of HL and J^2 were close to $\chi^2(G-2)$ when the model contained only a continuous covariate, but when the linear predictor contained one continuous and a dichotomous term, or two continuous covariates and an interaction term, the results indicated that the degrees of freedom of the distributions of HL and J^2 may be reduced. We observed similar results in our null simulations, which were conducted with fewer replications, but over a greater variety of settings. Overall, T_G maintained the Type I error rate well under all of the links studied, regardless of the number or types of terms included in the model. On the other hand, the Type I error rates of HL and J^2 were more dependent on the characteristics of the linear predictor. They were generally well maintained when the model setting contained only a single continuous covariate, but were often lower than expected when the model was more complex, especially if the one of the terms was dichotomous, quadratic, or an interaction term.

All three of the statistics had similar power to detect the omission of a term from the linear predictor in the fitted model. Power ranged from low to high, with the greatest power observed when the omitted term was a quadratic or interaction term. The three statistics had low to moderate power to detect an incorrectly specified probit, log-log and complementary log-log link function when the true underling model was logistic. The highest power was observed under the log and log-log links. When evaluating incorrectly fit log binomial models, T_{LB} had more power than HL and J^2 to detect an incorrectly specified log link, especially when the underlying logistic model included dichotomous, quadratic or interaction terms.

6.5 Limitations of This Research

A limitation of any simulation study is that it is impossible to investigate all conceivable scenarios. This is due to limitations in time and funding, and the need to state a finite focus for the thesis. We limited our focus in several ways. First we compared only three goodness-of-fit

statistics, HL, J^2, and $T_{\mathcal{G}}$ (including the original T). We selected these because of their apparent close algebraic relationships, and because little published work had been done to compare their performances under the same grouping method.

Secondly, we chose to compare the statistics under only one grouping method, the deciles-of-risk method. More grouping methods were not considered, in part due to time constraints. Although the deciles-of-risk method has some deficiencies, it is a well-established method cited widely in the literature, and it is easily calculated and intuitively appealing to applied scientists.

We also limited the number of settings studied, again due mainly to restrictions on time. Although the number and variability of the settings were limited, we felt that there was enough variation in the combinations studied to be able to make some assessment of the performances of the statistics.

A further limitation is that only four non-canonical links were studied. This again was mainly dictated by time constraints. We strove to evaluate the statistics under some of the more commonly used GLM for Bernoulli outcomes, focusing on the logistic (as the original T), and the non-canonical probit, log-log, complimentary log-log, and log binomial models. Other possible links that could be studied are the log complement and the identity links (see Hardin, et al. (2007)).

Another limitation is that we were unable to recommend an adjustment for the degrees-of-freedom of HL and J^2. Because the distribution seemed to vary depending on the types of covariates in the model and their relative influence in the model, it was not possible to determine a single adjustment.

6.6 Future Research

The scope of this research is finite, and has produced some further questions that could be pursued in future research. More work needs to be done to establish the distributional properties of J^2. Specifically, further research needs to be done to determine its distribution under

different grouping methods. This should include simulation studies where J^2 is calculated using grouping strategies that partition the covariate space, as well as methods that reference the observed data.

In addition, simulation studies using more complex settings (i.e. >3 covariates) could be conducted to further establish the distributional characteristics and performance of HL, J^2, and $T_{\mathcal{G}}$ (including T).

Another potential research direction would be to study how HL, J^2, and $T_{\mathcal{G}}$ can be implemented when data are sparse. One approach is to use parametric Monte Carlo bootstrap methods to generate empirical distributions that represent the distributions of HL, J^2, and $T_{\mathcal{G}}$ when the null hypothesis holds. The methods used by Tollenaar and Mooijaart (2003), who performed a simulation study on several well-known goodness-of-fit tests whose distributions are affected by sparse data, could be adopted for this research.

Finally, another possible research direction would be to apply the $T_{\mathcal{G}}$ goodness-of-fit statistic to the evaluation of other non-canonical GLM. This might include simulation studies of other non-canonical GLM with Bernoulli outcomes, such as the log complement link (Hardin, et al. 2007). Another tack would be to extend the $T_{\mathcal{G}}$ methods to non-canonical GLM with outcomes from the Poisson distribution.

Appendix A Derivation of Terms for the Calculation of $T_{\mathcal{G}}$

The derivation of the terms need for the calculation of $T_{\mathcal{G}}$ when assessing logit, probit, log-log, complementary log-log, and log binomial models are given below.

A1 Canonical Logit Link

Under the logistic model the additional term (5.11), using (2.29), is

$$h(\mathbf{x}'\boldsymbol{\beta}) = \frac{\partial \eta_0}{\partial \zeta_0} = 1 \tag{A1.1}$$

Thus the model becomes the original Tsiatis model, (3.9), and the usual Tsiatis statistic (3.11), T, is used to test the fit of the null model to the data.

A2 Non-Canonical Links

A2.1 Probit Link ($\mathbf{T_{Pr}}$)

Under the probit model, the term (5.11), using (2.34), is

$$h(\mathbf{x}'\boldsymbol{\beta}) = \frac{\partial \eta_0}{\partial \zeta_0}$$

$$= \frac{\theta_{\Pr}(\mathbf{x})\{1-\theta_{\Pr}(\mathbf{x})\}}{\phi(\eta_0)} \tag{A1.2}$$

and thus the model (5.10) is

$$\theta_{\Pr}(\mathbf{x},\mathbf{I}) = \Phi\{\eta\} \tag{A1.3}$$

where

$$\eta = \mathbf{x}'\boldsymbol{\beta} + \frac{\theta_{\Pr}(\mathbf{x})\{1-\theta_{\Pr}(\mathbf{x})\}}{\phi(\eta_0)} \sum_{g=1}^{G} \gamma_g I^{(g)} \tag{A1.4}$$

The derivatives necessary for calculating the terms of T_{Pr} are

$$\frac{\partial \theta_{\mathrm{Pr}}(\mathbf{x},\mathbf{I})}{\partial \eta} = \frac{\partial \Phi\{\eta\}}{\partial \eta}$$

$$= \phi(\eta) \tag{A1.5}$$

$$\frac{\partial h(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_k} = \frac{\partial h(\mathbf{x}'\boldsymbol{\beta})}{\partial \theta_{\mathrm{Pr}}(\mathbf{x})} \frac{\partial \theta_{\mathrm{Pr}}(\mathbf{x})}{\partial \eta_0} \frac{\partial \eta_0}{\partial \beta_k}$$

$$= \frac{\partial \left[\dfrac{\theta_{\mathrm{Pr}}(\mathbf{x})\{1-\theta_{\mathrm{Pr}}(\mathbf{x})\}}{\phi(\eta_0)} \right]}{\partial \theta_{LL}(\mathbf{x})} \frac{\partial \theta_{\mathrm{Pr}}(\mathbf{x})}{\partial \eta_0} \frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial \beta_k}$$

$$= \left\{ \frac{\{1-2\theta_{\mathrm{Pr}}(\mathbf{x})\}\phi(\mathbf{x}) - (-\eta_0)\theta_{\mathrm{Pr}}(\mathbf{x})\{1-\theta_{\mathrm{Pr}}(\mathbf{x})\}}{\{\phi(\eta_0)\}^2} \right\} \phi(\eta_0) x_k$$

$$= \frac{\{1-2\theta_{\mathrm{Pr}}(\mathbf{x})\}\phi(\mathbf{x}) + \eta_0 \theta_{\mathrm{Pr}}(\mathbf{x})\{1-\theta_{\mathrm{Pr}}(\mathbf{x})\}}{\phi(\eta_0)} x_k \tag{A1.6}$$

since

$$\frac{\partial \left[\dfrac{\theta_{\mathrm{Pr}}(\mathbf{x})\{1-\theta_{\mathrm{Pr}}(\mathbf{x})\}}{\phi(\eta_0)} \right]}{\partial \theta_{\mathrm{Pr}}(\mathbf{x})} = \frac{\dfrac{\partial \left[\theta_{\mathrm{Pr}}(\mathbf{x})\{1-\theta_{\mathrm{Pr}}(\mathbf{x})\} \right]}{\partial \theta_{\mathrm{Pr}}(\mathbf{x})} \phi(\eta_0) - \dfrac{\partial \left[\phi(\eta_0) \right]}{\partial \theta_{\mathrm{Pr}}(\mathbf{x})} \theta_{\mathrm{Pr}}(\mathbf{x})\{1-\theta_{\mathrm{Pr}}(\mathbf{x})\}}{\{\phi(\eta_0)\}^2}$$

$$= \frac{\{1-2\theta_{\mathrm{Pr}}(\mathbf{x})\}\phi(\mathbf{x}) - (-\eta_0)\theta_{\mathrm{Pr}}(\mathbf{x})\{1-\theta_{\mathrm{Pr}}(\mathbf{x})\}}{\{\phi(\eta_0)\}^2} \tag{A1.7}$$

and

$$\frac{\partial \phi(\eta_0)}{\partial \theta_{\mathrm{Pr}}(\mathbf{x})} = \frac{\partial}{\partial \theta_{\mathrm{Pr}}(\mathbf{x})} \frac{\partial \theta_{\mathrm{Pr}}(\mathbf{x})}{\partial \eta_0}$$

$$= \frac{\partial}{\partial \theta_{\mathrm{Pr}}(\mathbf{x})} \frac{\partial \eta_0}{\partial \eta_0} \frac{\partial \theta_{\mathrm{Pr}}(\mathbf{x})}{\partial \eta_0}$$

$$= \frac{\partial \eta_0}{\partial \theta_{\mathrm{Pr}}(\mathbf{x})} \frac{\partial^2 \theta_{\mathrm{Pr}}(\mathbf{x})}{\partial \eta_0 \partial \eta_0}$$

$$= \left\{ \frac{1}{\phi(\eta_0)} \right\} \{ -\eta_0 \phi(\eta_0) \}$$

$$= -\eta_0 \qquad \text{(A1.8)}$$

(Hardin, et al. 2007)

Under the null hypothesis, $\hat{\theta}_{\mathrm{Pr}}(\mathbf{x}, \mathbf{I}) = \hat{\theta}_{\mathrm{Pr}}(\mathbf{x})$ and $\hat{\phi}(\eta) = \hat{\phi}(\eta_0)$, and will be denoted generally as $\hat{\theta}_{\mathrm{Pr}}$ and $\hat{\phi}$ respectively. Thus under the null hypothesis the gth term of the score vector, (5.21), becomes

$$\left. \frac{\partial l}{\partial \gamma_g} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \boldsymbol{\gamma}=0} = \sum_{i=1}^{n} \left(y_i - \hat{\theta}_{\mathrm{Pr}_i} \right) I_i^{(g)} \qquad \text{(A1.9)}$$

The elements of the covariance matrix $\mathbf{V}$ under the null hypothesis are

$$A_{gg'}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \boldsymbol{\gamma}=0} = \sum_{i=1}^{n} \hat{\theta}_{\mathrm{Pr}} \left(1 - \hat{\theta}_{\mathrm{Pr}} \right) I_i^{(g)} I_i^{(g')} \qquad \text{(A1.10)}$$

$$B_{gk}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \boldsymbol{\gamma}=0} = \sum\nolimits_{i=1}^{n} \left(\frac{\partial \hat{\theta}_{\mathrm{Pr}}}{\partial \eta_0} \right)_i^2 I_i^{(g)} x_{ik}$$

$$= \sum_{i=1}^{n} \hat{\phi}_i I_i^{(g)} x_{ik} \qquad \text{(A1.11)}$$

$$C_{kk'}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \boldsymbol{\gamma}=0} = \sum\nolimits_{i=1}^{n} \left(\frac{\partial \hat{\theta}_{\mathrm{Pr}}}{\partial \eta_0} \right)_i^2 \left(\frac{\partial \zeta_0}{\partial \hat{\theta}_{\mathrm{Pr}}} \right)_i x_{ik} x_{ik'}$$

$$= \sum_{i=1}^{n} \frac{\hat{\phi}_i^2}{\hat{\theta}_{\mathrm{Pr}_i} \left(1 - \hat{\theta}_{\mathrm{Pr}_i} \right)} x_{ik} x_{ik'} \qquad \text{(A1.12)}$$

A2.2 Log-log Link (TLL)

Under the log-log model the term (5.11), using (2.41), is

$$h(\mathbf{x}'\boldsymbol{\beta}) = \frac{\partial \eta_0}{\partial \zeta_0}$$

$$= \frac{1-\theta_{LL}(\mathbf{x})}{-\ln \theta_{LL}(\mathbf{x})} \tag{A1.13}$$

and thus the model (5.10) is

$$\theta_{LL}(\mathbf{x},\mathbf{I}) = \exp\left[-\exp\{-\eta\}\right] \tag{A1.14}$$

where

$$\eta = \mathbf{x}'\boldsymbol{\beta} + \frac{1-\theta_{LL}(\mathbf{x})}{-\ln \theta_{LL}(\mathbf{x})} \sum_{g=1}^{G} \gamma_g I^{(g)} \tag{A1.15}$$

The derivatives necessary for calculating the terms of T_{LL} are

$$\frac{\partial \theta_{LL}(\mathbf{x},\mathbf{I})}{\partial \eta} = \frac{\partial \exp\left(-\exp(-\eta)\right)}{\partial \eta}$$

$$= -\theta_{LL}(\mathbf{x},\mathbf{I}) \ln \theta_{LL}(\mathbf{x},\mathbf{I}) \tag{A1.16}$$

$$\frac{\partial h(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_k} = \frac{\partial h(\mathbf{x}'\boldsymbol{\beta})}{\partial \theta_{LL}(\mathbf{x})} \frac{\partial \theta_{LL}(\mathbf{x})}{\partial \eta_0} \frac{\partial \eta_0}{\partial \beta_k}$$

$$= \frac{\left[\partial \left\{ \dfrac{1-\theta_{LL}(\mathbf{x})}{-\ln \theta_{LL}(\mathbf{x})} \right\}\right]}{\partial \theta_{LL}(\mathbf{x})} \frac{\partial \left\{\exp\left(-\exp\left(-\eta_0\right)\right)\right\}}{\partial \eta_0} \frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial \beta_k}$$

$$= \frac{\theta_{LL}(\mathbf{x}) \ln \theta_{LL}(\mathbf{x}) + 1 - \theta_{LL}(\mathbf{x})}{\left\{\ln \theta_{LL}(\mathbf{x})\right\}^2 \theta_{LL}(\mathbf{x})} \left\{-\ln \theta_{LL}(\mathbf{x})\right\} \theta_{LL}(\mathbf{x}) x_k$$

$$= \frac{\theta_{LL}(\mathbf{x}) \ln \theta_{LL}(\mathbf{x}) + 1 - \theta_{LL}(\mathbf{x})}{-\ln \theta_{LL}(\mathbf{x})} x_k \tag{A1.17}$$

Under the null hypothesis, $\hat{\theta}_{LL}(\mathbf{x},\mathbf{I}) = \hat{\theta}_{LL}(\mathbf{x})$, and will be denoted generally as $\hat{\theta}_{LL}$. Thus under the null hypothesis the *g*th term of the score vector, (5.21), becomes

$$\left.\frac{\partial l}{\partial \gamma_g}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum_{i=1}^{n}\left\{y_i - \hat{\theta}_{LL_i}\right\} I_i^{(g)} \tag{A1.18}$$

The elements of the covariance matrix $\mathbf{V}$ under the null hypothesis are

$$A_{gg'}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum_{i=1}^{n}\hat{\theta}_{LL_i}\left(1-\hat{\theta}_{LL_i}\right) I_i^{(g)} I_i^{(g')} \tag{A1.19}$$

$$B_{gk}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum\nolimits_{i=1}^{n}\left(\frac{\partial\hat{\theta}_{LL}}{\partial\eta_0}\right)_i^2 I_i^{(g)} x_{ik}$$

$$= \sum\nolimits_{i=1}^{n} -\hat{\theta}_{LL_i}\ln\hat{\theta}_{LL_i} I_i^{(g)} x_{ik} \tag{A1.20}$$

$$C_{kk'}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum\nolimits_{i=1}^{n}\left(\frac{\partial\hat{\theta}_{LL}}{\partial\eta_0}\right)_i^2\left(\frac{\partial\zeta_0}{\partial\hat{\theta}_{LL}}\right)_i x_{ik}x_{ik'}$$

$$= \sum\nolimits_{i=1}^{n}\frac{\hat{\theta}_{LL_i}}{1-\hat{\theta}_{LL_i}}\left(\ln\hat{\theta}_{LL_i}\right)^2 x_{ik}x_{ik'} \tag{A1.21}$$

A2.3 Complementary Log-log Link ($\mathbf{T_{Cll}}$)

Under the complementary log-log model the additional term (5.11), using (2.42), is

$$h\left(\mathbf{x}'\boldsymbol{\beta}\right) = \frac{\partial\eta_0}{\partial\zeta_0}$$

$$= \frac{\theta_{Cll}\left(\mathbf{x}\right)}{-\ln\left\{1-\theta_{Cll}\left(\mathbf{x}\right)\right\}} \tag{A1.22}$$

and thus the model (5.10) is

$$\theta_{Cll}\left(\mathbf{x},\mathbf{I}\right) = 1-\exp\left(-\exp\left(\eta\right)\right) \tag{A1.23}$$

where

$$\eta = \mathbf{x}'\boldsymbol{\beta} + \frac{\theta_{Cll}\left(\mathbf{x}\right)}{-\ln\left\{1-\theta_{Cll}\left(\mathbf{x}\right)\right\}}\sum_{g=1}^{G}\gamma_g I^{(g)} \tag{A1.24}$$

The derivative necessary for calculating the terms is

$$\frac{\partial \theta_{Cll}(\mathbf{x},\mathbf{I})}{\partial \eta} = \frac{\partial\left[1-\exp\{-\exp(\eta)\}\right]}{\partial \eta}$$

$$= -\{1-\theta_{Cll}(\mathbf{x},\mathbf{I})\}\ln\{1-\theta_{Cll}(\mathbf{x})\} \qquad \text{(A1.25)}$$

$$\frac{\partial h(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_k} = \frac{\partial h(\mathbf{x}'\boldsymbol{\beta})}{\partial \theta_{Cll}(\mathbf{x})}\frac{\partial \theta_{Cll}(\mathbf{x})}{\partial \eta_0}\frac{\partial \eta_0}{\partial \beta_k}$$

$$= \frac{\partial\left[\dfrac{\theta_{Cll}(\mathbf{x})}{-\ln\{1-\theta_{Cll}(\mathbf{x})\}}\right]}{\partial \theta_{Cll}(\mathbf{x})}\frac{\partial\{1-\exp(-\exp(\eta_0))\}}{\partial \eta_0}\frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial \beta_k}$$

$$= \frac{-\ln\{1-\theta_{Cll}(\mathbf{x})\}+\theta_{Cll}(\mathbf{x})\ln\{1-\theta_{Cll}(\mathbf{x})\}-\theta_{Cll}(\mathbf{x})}{\left(\ln\{1-\theta_{Cll}(\mathbf{x})\}\right)^2\{1-\theta_{Cll}(\mathbf{x})\}} \left[-\{1-\theta_{Cll}(\mathbf{x},\mathbf{I})\}\ln\{1-\theta_{Cll}(\mathbf{x})\}\right]x_k$$

$$= \frac{-\ln\{1-\theta_{Cll}(\mathbf{x})\}+\theta_{Cll}(\mathbf{x})\ln\{1-\theta_{Cll}(\mathbf{x})\}-\theta_{Cll}(\mathbf{x})}{-\ln\{1-\theta_{Cll}(\mathbf{x})\}}x_k \qquad \text{(A1.26)}$$

Under the null hypothesis, $\hat{\theta}_{Cll}(\mathbf{x},\mathbf{I}) = \hat{\theta}_{Cll}(\mathbf{x})$, and will be denoted generally as $\hat{\theta}_{Cll}$. Under the null hypothesis the gth term of the score vector, (5.21), becomes

$$\left.\frac{\partial l}{\partial \gamma_g}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum_{i=1}^{n}\left(y_i - \hat{\theta}_{Cll_i}\right)I_i^{(g)} \qquad \text{(A1.27)}$$

The elements of the covariance matrix $\mathbf{V}$ under the null hypothesis are

$$A_{gg'}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum_{i=1}^{n}\hat{\theta}_{Cll}\left(1-\hat{\theta}_{Cll}\right)I_i^{(g)}I_i^{(g')} \qquad \text{(A1.28)}$$

$$B_{gk}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum\nolimits_{i=1}^{n}\left(\frac{\partial \hat{\theta}_{Cll}}{\partial \eta_0}\right)_i^2 I_i^{(g)}x_{ik}$$

$$= \sum_{i=1}^{n} -\left(1-\hat{\theta}_{Cll_i}\right)\ln\left(1-\hat{\theta}_{Cll_i}\right)I_i^{(g)}x_{ik} \qquad \text{(A1.29)}$$

$$C_{kk'}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\,\gamma=0} = \sum_{i=1}^{n}\left(\frac{\partial\hat{\theta}_{Cll}}{\partial\eta_0}\right)_i^2\left(\frac{\partial\zeta_0}{\partial\hat{\theta}_{Cll}}\right)_i x_{ik}x_{ik'}$$

$$= \sum_{i=1}^{n}\frac{1-\hat{\theta}_{Cll_i}}{\hat{\theta}_{Cll_i}}\left\{\ln\left(1-\hat{\theta}_{Cll_i}\right)\right\}^2 x_{ik}x_{ik'} \qquad \text{(A1.30)}$$

A2.4 Log Link (T_{LB})

Under the log binomial model, which has the log link function, the additional term (5.11), using (2.46), is

$$h\left(\mathbf{x}'\boldsymbol{\beta}\right) = \frac{\partial\eta_0}{\partial\zeta_0}$$

$$= 1-\theta_{LB}\left(\mathbf{x}\right) \qquad \text{(A1.31)}$$

and thus the model (5.10) is

$$\theta_{LB}\left(\mathbf{x},\mathbf{I}\right) = \exp\left(\eta\right) \qquad \text{(A1.32)}$$

where

$$\eta = \mathbf{x}'\boldsymbol{\beta} + \left\{1-\theta_{LB}\left(\mathbf{x}\right)\right\}\sum_{g=1}^{G}\gamma_g I^{(g)} \qquad \text{(A1.33)}$$

The derivatives necessary for calculating the terms of T_{LB} are

$$\frac{\partial\theta_{LB}\left(\mathbf{x},\mathbf{I}\right)}{\partial\eta} = \frac{\partial\exp\left(\eta\right)}{\partial\eta}$$

$$= \theta_{LB}\left(\mathbf{x},\mathbf{I}\right) \qquad \text{(A1.34)}$$

$$\frac{\partial h\left(\mathbf{x}'\boldsymbol{\beta}\right)}{\partial\beta_k} = \frac{\partial h\left(\mathbf{x}'\boldsymbol{\beta}\right)}{\partial\theta_{Cll}\left(\mathbf{x}\right)}\frac{\partial\theta_{Cll}\left(\mathbf{x}\right)}{\partial\eta_0}\frac{\partial\eta_0}{\partial\beta_k}$$

$$= \frac{\partial\left\{1-\theta_{LB}\left(\mathbf{x}\right)\right\}}{\partial\theta_{LB}\left(\mathbf{x}\right)}\frac{\partial\left\{\exp\left(\eta_0\right)\right\}}{\partial\eta_0}\frac{\partial\mathbf{x}'\boldsymbol{\beta}}{\partial\beta_k}$$

$$= -\theta_{LB}\left(\mathbf{x}\right)x_k \qquad \text{(A1.35)}$$

Under the null hypothesis, $\hat{\theta}_{LB}(\mathbf{x},\mathbf{I}) = \hat{\theta}_{LB}(\mathbf{x})$, and will be denoted generally as $\hat{\theta}_{LB}$.

Thus under the null hypothesis the gth term of the score vector, (5.21), becomes

$$\left.\frac{\partial l}{\partial \gamma_g}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum_{i=1}^{n} y_i - \hat{\theta}_{LB_i} I_i^{(g)} \tag{A1.36}$$

The elements of the covariance matrix $\mathbf{V}$ under the null hypothesis are

$$A_{gg'}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum_{i=1}^{n} \hat{\theta}_{LB_i}\left(1-\hat{\theta}_{LB_i}\right) I_i^{(g)} I_i^{(g')} \tag{A1.37}$$

$$B_{gk}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum_{i=1}^{n} \left(\frac{\partial \hat{\theta}_{LB}}{\partial \eta_0}\right)_i^2 I_i^{(g)} x_{ik}$$

$$= \sum_{i=1}^{n} \hat{\theta}_{LB_i} I_i^{(g)} x_{ik} \tag{A1.38}$$

$$C_{kk'}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\ \boldsymbol{\gamma}=0} = \sum_{i=1}^{n} \left(\frac{\partial \hat{\theta}_{LB}}{\partial \eta_0}\right)_i^2 \left(\frac{\partial \zeta_0}{\partial \hat{\theta}_{LB}}\right)_i x_{ik} x_{ik'}$$

$$= \sum_{i=1}^{n} \frac{\hat{\theta}_{LB_i}}{1-\hat{\theta}_{LB_i}} x_{ik} x_{ik'} \tag{A1.39}$$

Bibliography

Agresti, A. (2007), ***An Introduction to Categorical Data Analysis*** **(second ed.), New York, New York: Wiley - Interscience.**

Andrews, D. W. (1988), "Chi-Square Diagnostic Tests for Econometric Models: Introduction and Applications," ***Journal of Econometrics*****, 37, 135-156.**

Archer, K. J., Lemeshow, S., and Hosmer, D. W. (2007), "Goodness-of-Fit Tests for Logistic Regression Models When Data Are Collected Using a Complex Sampling Design," ***Computational Statistics & Data Analysis*****, 51, 4450-4464.**

Azzalini, A., Bowman, A. W., and Härdle, W. (1989), "On the Use of Nonparametric Regression for Model Checking," ***Biometrika*****, 76, 1-11.**

Barnhart, H. X., and Williamson, J. M. (1998), "Goodness-of-Fit Tests for G.E.E. Modeling with Binary Responses," ***Biometrics*****, 720-729.**

Bertolini, G., Damico, R., Nardi, D., Tinazzi, A., and Apolone, G. (2000), "One Model, Several Results: The Paradox of the Hosmer-Lemeshow Goodness-of-Fit Test for the Logistic Regression Model," ***Journal of Epidemiology and Biostatistics*****, 5, 251-253.**

Blizzard, L., and Hosmer, D. W. (2006), "Parameter Estimation and Goodness-of-Fit in Log Binomial Regression," ***Biometrical Journal*****, 48, 5-22.**

Blizzard, L., and Hosmer, D. W. (2007), "The Log Multinomial Regression Model for Nominal Outcomes with More Than Two Attributes," ***Biometrical Journal*****, 49, 889-902.**

Bull, S. (1994), "Analysis of Attitudes toward Workplace Smoking Restrictions," ***Case Studies in Biometry*****, 249-271.**

Chernoff, H., and Lehmann, E. L. (1954), "The Use of Maximum Likelihood Estimates in X2 Tests for Goodness of Fit," ***The Annals of Mathematical Statistics*****, 25, 579-586.**

Copas, J. (1983), "Plotting P against X," ***Applied Statistics*****, 25-31.**

Copas, J. (1989), "Unweighted Sum of Squares Test for Proportions," ***Applied Statistics*****, 71-80.**

Cressie, N., and Read, T. R. (1984), "Multinomial Goodness-of-Fit Tests," ***Journal of the Royal Statistical Society. Series B (Methodological)*****, 440-464.**

Czado, C., and Munk, A. (2000), "Noncanonical Links in Generalized Linear Models– When Is the Effort Justified?," ***Journal of Statistical Planning and Inference*****, 87, 317-345.**

Deng, D. (2001), "*Goodness-of-fit, Score Test, Zero-Inflation and over-Dispersion in Generalized Linear Models,*" University of Windsor, Mathematics and Statistics.

Deng, D., and Paul, S. R. (2000), "Score Tests for Zero Inflation in Generalized Linear Models," *Canadian Journal of Statistics*, 28, 563-570.

Dreiseitl, S., and Osl, M. (2012), "Effects of Data Grouping on Calibration Measures of Classifiers," *Computer Aided Systems Theory–EUROCAST 2011*, 359-366.

Durst, M. (1979), "Personal Communication - Unpublished Thesis," 1043-1069.

Farrington, C. (1995), "Pearson Statistics, Goodness-of-Fit, and Overdispersion in Generalised Linear Models," in *Statistical Modelling*, ed. G. U. H. Seeber, Springer, pp. 109-116.

Farrington, C. (1996), "On Assessing Goodness of Fit of Generalized Linear Models to Sparse Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 349-360.

Graybill, F. A. (1976), *Theory and Application of the Linear Model*, Duxbury Press.

Halteman, W. A. (1980), "*A Goodness-of-Fit Statistic for Binary Logistic Regression*," University of Washington, Biomathematics.

Hardin, J. W., and Hilbe, J. M. (2007), *Generalized Linear Models and Extensions*, College Station, Texas: Stata Press.

Henderson, H. V., and Searle, S. R. (1981), "On Deriving the Inverse of a Sum of Matrices," *Siam Review*, 23, 53-60.

Horton, N. J., et al. (1999), "Goodness-of-Fit for G.E.E.: An Example with Mental Health Service Utilization," *Statistics in Medicine*, 18, 213-222.

Hosmer, D. W., and Hjort, N. L. (2002), "Goodness-of-Fit Processes for Logistic Regression: Simulation Results," *Statistics in Medicine*, 21, 2723-2738.

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997), "A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model," *Statistics in Medicine*, 16, 965-980.

Hosmer, D. W., and Lemeshow, S. (1980), "Goodness of Fit Tests for the Multiple Logistic Regression Model," *Communications in Statistics - Theory and Methods*, 9, 1043-1069.

Hosmer, D. W., and Lemeshow, S. (2000), *Applied Logistic Regression*, New York: Wiley.

Hosmer, D. W., Lemeshow, S., and Klar, J. (1988), "Goodness-of-Fit Testing for the Logistic Regression Model When the Estimated Probabilities Are Small," *Biometrical Journal*, 30, 911-924.

Kendall, M. G., Stuart, A., and Ord, J. K. (1994), *Kendall's Advanced Theory of Statistics* (Vol. 1 Distributional Theory, 6th ed.), London, New York: Edward Arnold, Halsted Press.

Kendall, M. G., Stuart, A., Ord, K., and Arnold, S. (1999), ***Kendall's Advanced Theory of Statistics*** **(Vol. 2A Classical Inference and the Linear Model 6ed.), ed. K. O. A. Stuart, S. Arnold London: Arnold.**

Kuss, O. (2002), "Global Goodness-of-Fit Tests in Logistic Regression with Sparse Data," ***Statistics in Medicine*****, 21, 3789-3801.**

le Cessie, S., and van Houwelingen, H. C. (1995), "Testing the Fit of a Regression Model Via Score Tests in Random Effects Models," ***Biometrics*****, 600-614.**

le Cessie, S., and van Houwelingen, J. C. (1991), "A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods," ***Biometrics*****, 1267-1282.**

Lemeshow, S., and Hosmer, D. W. (1982), "A Review of Goodness-of-Fit Statistics for Use in the Development of Logistic Regression Models," ***American Journal of Epidemiology*****, 115, 92-106.**

Lin, D. Y., Wei, L. J., and Ying, Z. (2002), "Model-Checking Techniques Based on Cumulative Residuals," ***Biometrics*****, 58, 1-12.**

Liu, Y., Nelson, P. I., and Yang, S.-S. (2012), "An Omnibus Lack of Fit Test in Logistic Regression with Sparse Data," ***Statistical Methods & Applications*****, 21, 437-452.**

McCullagh, P. (1985), "On the Asymptotic Distribution of Pearson's Statistic in Linear Exponential-Family Models," ***International Statistical Review/Revue Internationale de Statistique*****, 61-67.**

McCullagh, P. (1986), "The Conditional Distribution of Goodness-of-Fit Statistics for Discrete Data," ***Journal of the American Statistical Association*****, 81, 104-107.**

McCullagh, P., and Nelder, J. A. (1989), ***Generalized Linear Models*****, London: Chapman & Hall/CRC.**

Moore, D. S. (1971), "A Chi-Square Statistic with Random Cell Boundaries," ***The Annals of Mathematical Statistics*****, 42, 147-156.**

Moore, D. S., and Spruill, M. C. (1975), "Unified Large-Sample Theory of General Chi-Squared Statistics for Tests of Fit," ***The Annals of Statistics*****, 599-616.**

Mukhopadhyay, N. (2000), ***Probability and Statistical Inference*** **(Vol. 162), CRC.**

Nadaraya, E. A. (1964), "On Estimating Regression," ***Theory of Probability & Its Applications*****, 9, 141-142.**

Nelder, J. A., and Wedderburn, R. W. (1972), "Generalized Linear Models," ***Journal of the Royal Statistical Society. Series A (General)*****, 370-384.**

Osius, G., and Rojek, D. (1992), "Normal Goodness-of-Fit Tests for Multinomial Models with Large Degrees of Freedom," ***Journal of the American Statistical Association*****, 87, 1145-1152.**

Paul, S., and Deng, D. (2011), "Assessing Goodness of Fit of Generalized Linear Models to Sparse Data Using Higher Order Moment Corrections," *Sankhyā: Series B*, 74, 195-210.

Paul, S. R., and Deng, D. (2002), "Goodness of Fit of Generalized Linear Models to Sparse Data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 323-333.

Pearson, K. (1900), "X. On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50, 157-175.

Pigeon, J. G., and Heyse, J., F. (1999a), "A Cautionary Note About Assessing the Fit of Logistic Regression Models," *Journal of Applied Statistics*, 26, 847.

Pigeon, J. G., and Heyse, J. F. (1999b), "An Improved Goodness of Fit Statistic for Probability Prediction Models," *Biometrical Journal*, 41, 71-82.

Pregibon, D. (1981), "Logistic Regression Diagnostics," *The Annals of Statistics*, 9, 705-724.

Pulkstenis, E., and Robinson, T. J. (2002), "Two Goodness-of-Fit Tests for Logistic Regression Models with Continuous Covariates," *Statistics in Medicine*, 21, 79-93.

Rao, C. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York:

Rao, C. R. (1948), "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation," in *Proceedings of the Cambridge Philosophical Society*, Cambridge Univ Press, pp. 50-57.

Rao, C. R. (2002), *Linear Statistical Inference and Its Applications* (2 ed.), New York, NY: Wiley.

Read, T. R., and Cressie, N. A. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*, New York: Springer-Verlag.

Royston, P. (1992), "The Use of Cusums and Other Techniques in Modelling Continuous Covariates in Logistic Regression," *Statistics in Medicine*, 11, 1115-1129.

Searle, S. R. (2006), *Matrix Algebra Useful for Statistics*, Hoboken, N.J.: Wiley-Interscience.

Smyth, G. K. (2003), "Pearson's Goodness of Fit Statistic as a Score Test Statistic," *Institute of Mathematical Statistics: Lecture Notes-Monograph Series, Statistics and science: a festschrift for Terry Speed*, 40, 115-126.

Stata Statistical Software: Release 10. StataCorp LP, College Station, TX

Stata Statistical Software: Release 12. StataCorp LP, College Station, TX.

Stukel, T. A. (1988), "Generalized Logistic Models," *Journal of the American Statistical Association*, 83, 426-431.

Su, J. Q., and Wei, L. J. (1991), "A Lack-of-Fit Test for the Mean Function in a Generalized Linear Model," *Journal of the American Statistical Association*, 86, 420-426.

Tollenaar, N., and Mooijaart, A. (2003), "Type I Errors and Power of the Parametric Bootstrap Goodness-of-Fit Test: Full and Limited Information," *British Journal of Mathematical and Statistical Psychology*, 56, 271-288.

Tsiatis, A. A. (1980), "A Note on a Goodness-of-Fit Test for the Logistic Regression Model," *Biometrika*, 67, 250-251.

Watson, G. (1957), "The X2 Goodness-of-Fit Test for Normal Distributions," *Biometrika*, 44, 336-348.

Watson, G. S. (1964), *Smooth Regression Analysis*, Baltimore, MD: Defense Technical Information Center.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica: Journal of the Econometric Society*, 50, 1-25.

Windmeijer, F. A. (1995), "Goodness-of-Fit Measures in Binary Choice Models 1," *Econometric Reviews*, 14, 101-116.

Xie, X.-J., Pendergast, J., and Clarke, W. (2008), "Increasing the Power: A Practical Approach to Goodness-of-Fit Test for Logistic Regression Models with Continuous Predictors," *Computational Statistics & Data Analysis*, 52, 2703-2713.

Xie, X. (2005), "*A Goodness-of-Fit Test for Logistic Regression Models with Continuous Predictors*," University of Iowa, Biostatistics.